Colliding Partons at the LHC

Hayden Hollenbeck

August 20, 2021

MSc in Theoretical Physics

The University of Edinburgh

2021

Abstract

Fragmentation functions (FFs) describe the non-perturbative effects of a parton fragmenting into a hadron. As with their counterparts parton distribution functions (PDFs), FFs are key to applying perturbative QCD to hadron scattering. An overview of FFs is described here, noting the similarities and differences shared with PDFs. Special attention is paid to bridging numerical, theoretical, and experimental calculations.

This paper presents the results of reweighting the unpolarized neutral pion fragmentation function set `MAPFF` against proton-proton collision data. Unlike previous fits of FFs against hadron-hadron collisions, `MAPFF` is parameterized by a neural-network. The FFs are represented by a Monte Carlo ensemble of replicas, which are assigned weights based upon how well individual replicas predict proton-proton cross sections to next to leading order (NLO) in perturbative QCD. Because the original `MAPFF` fit, based on SIA and SIDIS data, accurately predicts the proton-proton cross sections, the reweighting does not significantly constrain the FF set. The successes and limitations of reweighting are discussed, and an unweighted FF set is reported. Future fits of other light hadron FFs will provide a validation of the techniques presented here.

Declaration

I declare that this dissertation was composed entirely by myself.

Chapter 1 introduces the subject and aims of the project, and describes previous work at a general level. It does not contain any original research.

Chapter 2 provides an introduction to the theoretical basis of parton distribution functions and fragmentation functions. Figures 2.8 and 2.10 were created by myself (with feedback from my supervisors) in order to demonstrate my understanding of computing theoretical cross sections. Libraries used include LHAPDF [1], APFEL [2], APFEL++ [3], and `yaml-cpp`. The figures were designed to replicate figures in [4] and [5] respectively. None of this chapter contains original research.

Chapters 3 and 4 describe the experimental data used in this paper and the methodology of reweighting applied to the data. They do not contain original research.

Chapter 5 details the results and analysis of reweighting. This chapter contains original work done entirely by myself. Theoretical cross sections were computed using code from [42] [43], which was supplied by post-doctoral researcher Emanuele Nocera. Filtering and some plotting scripts used the standard NumPy package. All other code was written by myself from scratch, with the exception of some filtering scripts written by Emanuele Nocera.

Appendices A and B contain auxiliary fit validations. The tests were not my original work, and their sources are cited in each section. The application of the tests to the reweighted fits is my own work.

Appendix C contains a link to a repository containing all code used in this project. A declaration of own work is supplied in the repository README file which identifies the code ownership.

All plots were created in Python using the standard `MatplotLib` package. Feynman diagrams were plotted from scratch in LaTeX's Tikz standard package.

Approximate word count: 11,218 words.

Personal Statement

I began this project by reading about the theory of fragmentation functions and techniques used to fit them. Concurrently, I started working on recreating the BABAR plots of SIA shown in the NNFF fitting paper [4]. After about 2 weeks I was able to reproduce the theoretical cross sections and uncertainties, as well as find collision data sets.

The next step was to attempt to recreate the SIDIS cross sections, which introduced a dependence upon parton distribution functions. Correctly understanding SIDIS is necessary to understand semi-inclusive hadron-hadron collisions, since they include a parton distribution function for each hadron. It took approximately another 2 weeks to recreate the general form of COMPASS data. Because the COMPASS observable was rather complicated and had a number of corrections, matching the general form was sufficient to move on to the project proper.

In July I started working on reweighting code, which was written entirely in C++. Code for unpolarized hadron-hadron cross sections was provided by Emanuele Nocera, as well as an example filtering script in Python. I tested the reweighting and unweighting code on some sample PHENIX data sets to make sure it had the desired results. After this was done, I identified all experimental data sets with semi-inclusive proton-proton collisions and computed their corresponding observables. While the observables were being calculated, I revised theory and experimental papers to verify parameters and techniques were correct. Near the end of July, the observables were finished and I began applying the reweighting code. During this time I implemented several tests to verify the results and determine the effects of various theoretical choices. Results were finalized throughout August.

During this time I kept in touch with my advisors Prof. Richard Ball and Dr. Emanuele Nocera via weekly email reports. I was able to resolve issues and make progress mainly independently, and when I couldn't, emails were able to clarify any issues. I met in person only twice (once at the beginning of June and once in the beginning of August) due to COVID social distancing rules.

I began writing methodology and theory sections in mid-July, and spent full time writing beginning the first week of August.

Acknowledgments

I would like to first and foremost thank my advisors Prof. Richard Ball and post-doctoral researcher Dr. Emanuele Nocera. This project was a continuation of their work, and only through their patient, consistent support was I able to attempt this project.

I thank my parents and family for their love and support. I miss you all greatly, and hope to be reunited soon.

I thank my roommates and friends I've met in Edinburgh this past year. Even in miserable times we've found a way to enjoy life.

And finally, I thank my friends back home who have stuck with me through thick and thin. Without them, I would never have made it to where I am today.

Contents

List of Tables

List of Figures

Chapter 1

Introduction

Hadronic scattering is governed by Quantum Chromodynamics (QCD), which describes the strong interaction that binds the internal hadron structure. Bound hadron states are non-perturbative, so observables dependent upon them cannot be computed ab initio in perturbation theory. However, at high energies QCD processes factorize into perturbative and non-perturbative components due to the asymptotic freedom of the strong coupling α_S [6]. The non-perturbative components consist of parton distribution functions (PDFs) and fragmentation functions (FFs) which describe the distribution of internal constituents of hadrons. Such constituent particles are known as "partons". Determination of these distribution functions tests the factorization and universality theorems of QCD. Accurate knowledge of distribution functions may also be used to predict results of strong processes, including hot QCD matter [7], as well as other distribution functions such as transverse momentum dependent (TMD) FFs [8].

A fitted distribution function consists of a method to calculate the central value and variance of any observable. The fitting may be done against any experimental data whose observable depends upon the distribution function. However, different processes are dominated by contributions from different distribution functions, e.g. proton-proton collisions are dominated by gluon FFs, while semi-inclusive annihilation is not [9]. By universality, determining the same distribution function from any two sets of observables will represent the same underlying "true" distribution [8]. Incorporating data from various processes therefore acts as a constraint, allowing it to more precisely reproduce the underlying exact distribution. Verifying that the new distribution function is indeed a constrained form of the previous distribution is thus a test of universality. Unless otherwise stated, universality will be assumed [8], and instead serve as way to validate fitting.

For a given hadron, PDFs describe the probability that one of its constituent partons participates in a collision, where the parton carries a fraction $x \in [0, 1]$ of the hadron's collinear momentum. In scattering events, the hadron appears in the initial state, such as in deep inelastic scattering, where a lepton interacts with a proton through an electroweak current. Because the non-perturbative behavior of the PDF is in the initial state, the final hadron states may be summed over, which is experimentally convenient. Thus there is significant data available to constraint PDF sets. Conversely, FFs describe the probability of a parton producing a given hadron with a momentum fraction z through hadronization.

In scattering events, FFs appear after the collision when a parton produces a hadron in the final state. Data used to constrain FFs must perform some type of particle identification of outgoing hadrons. Due to the experimental difficulty of particle identification, there's less available data to constrain FFs. The limited data demands careful treatment of uncertainties to ensure the best fit.

Initial attempts at fitting FFs focused on matching boundary behavior at $z \to 0$ and $z \to 1$ [8]. Intermediate behavior was fitted via an appropriate interpolation function between power scaling at the boundaries [10]. The fit is then limited by how accurately the form of the interpolation function was guessed, while any inaccuracies of the form are propagated into the uncertainty of the fit. Alternatively, FFs may be approximated by more general methods with no knowledge of their functional form. Following techniques of the NNPDF Collaboration [11], FFs have been approximated by training feed-forward neural networks [4] [5]. While the functional form of a neural network is too complicated to manipulate algebraically, it is straightforward to implement numerically, and assuming continuity, is guaranteed to reproduce the FF's functional form [12].

Fitting experimental data requires identifying the variance of the fit from experimental uncertainty and appropriately propagating the variance into observables. Explicit functional forms of the FFs allow for analytic descriptions of uncertainty, such as through individual parameters via the Hessian method [13]. For neural networks, the parameter count is too large to employ analytic techniques. Instead, the Monte Carlo (MC) technique is favored, where a statistical ensemble of MC replicas is generated preserving the average value and error of the experimental data [14]. A fit is performed against every MC replica individually, creating an ensemble of FFs. As in statistical mechanics, central values and variances of observables may be calculated via ensemble averages and moments of the distribution functions. For an initial fit, the probability that an arbitrary replica matches the underlying physical FF is uniform. This manifests in the ensemble average where all replicas are equally weighted. Introducing new experimental data to the fit alters the probability that a replica might match the total observed data set. Updating the probability with respect to new data is equivalent to adjusting the ensemble weights. This process is known as reweighting [15], which has been previously performed on PDFs. Reweighting is only dependent upon the MC replica ensemble of a FF and the new data to be fitted; it is agnostic of the fitting technique used to generate the original ensemble.

In this paper, reweighting was used to incorporate semi-inclusive proton-proton (pp) collisions into the integrated fragmentation function of the pion. The FF set used was `MAPFF10NLOPI`, which was fitted via a feed-forward neural network. This FF set was previously fitted against semi-inclusive annihilation (SIA) data and semi-inclusive deep inelastic scattering (SIDIS) data. The MAP collaboration has used neural networks to fit unidentified charged hadrons FFs against hadron-hadron collisions [16], however this was the first time a neural network FF for pions has been fitted against semi-inclusive hadron-hadron collisions.

Previous analysis of FFs against hadron-hadron collisions resulted in significant over prediction of observables [9]. However, the previous predictions used functional forms with limited (if available) uncertainty predictions which reduced the fit quality. The data-theory discrepancy was posited by the author to be due to the FF sets used, specifically

in the hardness of gluons in the high-z range at different energies. Results of reweighting the `MAPFF` set showed unbiased fitting is capable of describing the hadron-hadron data, as well as provide the uncertainties of the predictions.

1.1 Terminology and Notation

Quantities dependent on PDFs and FFs will often appear as sums over all possible partons. The parton flavor basis is denoted:

$$\{q, \overline{q}, g\} \qquad \text{where} \qquad q = u, d, s, c, b, t \tag{1.1}$$

Anti-quarks are denoted by $\overline{q}$.

Parton distribution functions of a hadron H into a parton i throughout this paper will be denoted as

$$f_i^H(x, \mu^2_{\text{fact}}), \qquad i = \{q, \overline{q}, g\} \tag{1.2}$$

where x is the momentum fraction and μ_{fact} is the factorization scale. In the case of a proton, $H = p$ is omitted by convention, i.e. the proton's PDFs are $\{f_i\}$.

Fragmentation functions of a parton i into a hadron H will be denoted as

$$D_i^H(z, \mu^2_{\text{frag}}), \qquad i = \{q, \overline{q}, g\} \tag{1.3}$$

where z is the momentum fraction and μ_{frag} is the fragmentation scale.

Both PDFs and FFs have similar properties and often follow the same path of analysis. In such cases, the generic name *distribution function* will be used. A distribution function associated to a parton i and hadron H is denoted

$$d_i^H(x, \mu^2), \qquad i = \{q, \overline{q}, g\} \tag{1.4}$$

Which corresponds to the same notation choices made for PDFs and FFs.

Chapter 2

Theory

This chapter introduces the concepts of parton distribution functions and fragmentation functions, and the primary methods to manipulate them. Theorems regarding the factorization of processes into distribution functions and their scaling dependence is briefly described, with emphasis on fragmentation functions. Three standard semi-inclusive process involving unpolarized fragmentation functions are described in greater detail at the end. Techniques for how to numerically calculate distribution functions and observables is emphasized throughout.

2.1 Introduction to Distribution Functions

The running of the strong coupling $\alpha_S(Q^2)$ disallows expansions in the coupling at low energies since the coupling grows large for small Q^2 [17]. Inelastically probing a strong process will therefore be non-perturbative, and produce a complicated bound state rather than an individual particle. Treatment of this bound state may be encompassed in the conjugate diagram as shown in figure 2.1. Applying the optical theorem by taking a vertical cut through the diagram gives the modulus squared of the desired original amplitude. Thus inelastic strong processes are naturally described via cut diagrams.

For a scattering event, such as in figure 2.2, the amplitude is dependent on the probing momentum q, which is conjugate to a spacetime interval ξ that separates the currents at

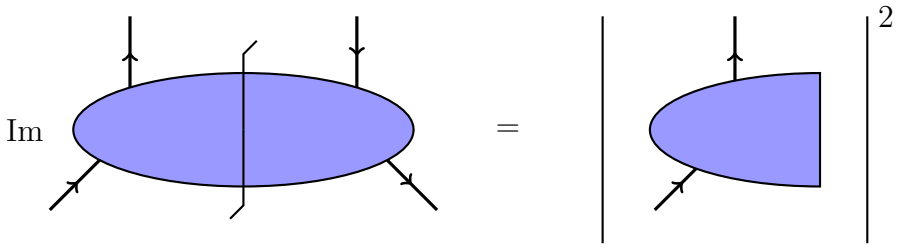

Figure 2.1: Application of the optical theorem to define cut diagrams (left).

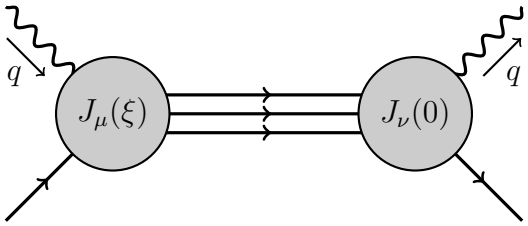

Figure 2.2: Forward hadronic current scattering process involving an electroweak current J_μ. Currents are separated by a spacetime interval ξ conjugate to the probing momentum q

two hadronic vertices [18]:

$$W_{\mu\nu} = \frac{1}{4M} \int \frac{\mathrm{d}^4\xi}{2\pi} e^{iq\cdot\xi} \langle p| \left[J_\mu(\xi), J_\nu(0)\right] |p\rangle \tag{2.1}$$

The dominant contributions to this amplitude occur at the singularities of the current commutator, which is when the interval ξ is light-like [18]. At such separations, both the time and space intervals vanish at high probing energies q^2, so the hadron being probed appears constant in time. Thus probing a hadron is time and space independent. This is the basis of the impulse approximation [18], which is the starting point for the quark parton model (QPM) [17].

In the QPM, the probing particles incoherently scatter off constituent particles of the parton, namely quarks and gluons. Processes factor, as shown in figure 2.3, into a hard cross section, involving only the partons and the probing particle, and a distribution function. The hard cross section may be directly computed using standard perturbation theory. The distribution contains all of the non-perturbative strong effects. For a diagram with an incoming hadron and outgoing parton, the distribution function is called a parton distribution function (PDF), while for an incoming parton and outgoing hadron the distribution function is called a fragmentation function (FF). Their similarities may be seen in figure 2.4.

In the QPM, PDFs are commonly defined with respect to deep inelastic scattering (DIS)

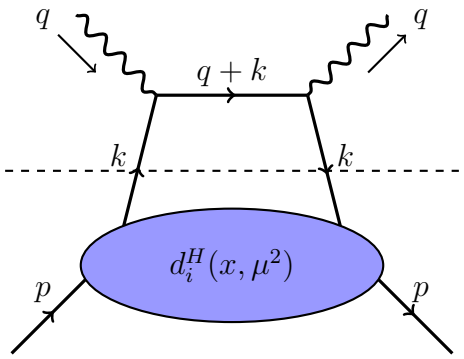

Figure 2.3: Handbag diagram of an electroweak current probing the constituent partons of a hadron. Separation of the hadronic and partonic components is shown by the dashed line.

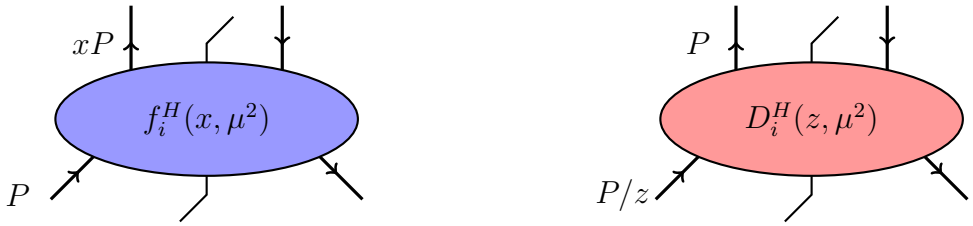

Figure 2.4: Defining cut diagrams of the PDF (left) and FF (right). P is the momentum of the hadron H.

[17] where a lepton scatters off of a hadron. An electroweak current probes the hadron via the handbag diagram 2.3, where the hadronic tensor may be rewritten as [18]:

$$W_{\mu\nu}(p,q) = \sum_{i=q,\bar{q},g} \int \mathrm{d}^4k\, B_i^H(p,q) w_{\mu\nu}^i(q,k) \delta[(k+q)^2] \tag{2.2}$$

Where i is a sum over partons, $w_{\mu\nu}^i$ is the hard cross-section involving parton i, and B_i^H is the unpolarized strong vertex. The delta function constrains the struck parton i to carry a fraction $x \in [0,1]$ of the hadron's momentum: $k = xp$. In DIS this is equivalent to the Bjorken x-variable [18]. PDFs q_i^H are then defined as [17]:

$$f_i^H(x) = \frac{\pi}{4} \int \mathrm{d}^2k_T\, B_i^H(p \cdot k) \tag{2.3}$$

where k_T^2 is the momentum transverse to q. In QPM, PDFs may be interpreted as the probability density of finding a parton i in the hadron H carrying a momentum xp. As with (2.2), observables are split into a partonic process and PDF, which are integrated over the full range of momentum fractions. A sum over all partons accounts for all possible partonic processes. This decomposition is explored in section 2.2.

Conversely, FFs in the QPM appear as the probability density of producing an outgoing hadron H from a parton i. The total unpolarized fragmentation function F^H may be defined with respect to semi-inclusive annihilation of leptons (figure 2.7) at a center of mass energy $\sqrt{s}$ [17]:

$$F^H(x,s) = \frac{1}{\sigma_{\text{tot}}} \frac{\mathrm{d}\sigma}{\mathrm{d}x}(l^+l^- \to HX) \tag{2.4}$$

where $x = 2E_H/\sqrt{s} \leq 1$ is the fraction of the total energy that H carries. F^H may be decomposed into the fragmentation functions D_i^H of individual partons i via [17]:

$$F^H(x,s) = \sum_{i=q,\bar{q},g} \int_x^1 \frac{\mathrm{d}z}{z} C_i(s;z,\alpha_S) D_i^H(x/z,s) \tag{2.5}$$

As with PDFs, the FFs carry the non-perturbative effects of the strong interaction. Cross sections involving FFs have a similar decomposition to PDFs, which is also explored in section 2.2.

Distribution functions obey certain symmetries under charge conjugation and isospin symmetry [8]. The charge conjugation symmetries relevant to this paper are:

$$d_q^{H^\pm} = d_{\bar{q}}^{H^\mp} \qquad \text{and} \qquad d_g^{H^\pm} = d_g^{H^\mp} \tag{2.6}$$

Under isospin symmetry, the charged pion distribution functions are related to the neutral pion:

$$\frac{1}{2}\left[d_i^{\pi^+} + d_i^{\pi^-}\right] = d_i^{\pi^0} \tag{2.7}$$

Up to a factor, the neutral pion and summed charged pion distribution functions are equal. Therefore, knowledge of the neutral pion cannot determine either of the charged pion distributions.

2.1.1 Mellin Transformations

Quantities involving distribution functions often appear as components of a Mellin convolution [3]:

$$(f \otimes g)(x) = \int_0^1 \mathrm{d}y \int_0^1 \mathrm{d}z\, f(y)g(z)\delta(x - yz) = \int_x^1 \frac{\mathrm{d}z}{z} f(x/z)g(z) \tag{2.8}$$

For example, equation (2.5) may be written as:

$$F^H(x, s) = \sum_{i=q,\bar{q},g} (C_i \otimes D_i^H)(x, s) \tag{2.9}$$

The factorization theorems (section 2.2) and DGLAP equations (section 2.3) will also be written in terms of Mellin convolutions. As with (Fourier) convolutions, there's a corresponding convolution theorem using the Mellin transformation:

$$\widetilde{f \otimes g}(n) = \tilde{f}(n) \cdot \tilde{g}(n) \tag{2.10}$$

where $\tilde{f}(n)$ is the Mellin transformation, defined alongside its inverse transformation [3]:

$$\tilde{f}(n) = \int_0^1 \mathrm{d}x\, x^{n-1} f(x) \qquad \text{and} \qquad f(x) = \frac{-1}{2\pi i} \int_{a-\infty}^{a+\infty} \mathrm{d}n\, x^{-n} \tilde{f}(n) \tag{2.11}$$

where a is chosen such that all poles of the integrand are to the left of the integration contour. These Mellin techniques are particularly useful in a direct analysis of distribution functions from field theory. Recall the forward scattering in figure 2.2, which admits the hadronic tensor $W_{\mu\nu}$ in equation (2.1). The currents are defined as local operators, so their commutator appears as a bilocal operator. Wilson's operator product expansion may be applied to the commutator to produce a series of bilocal operators $\mathcal{O}_n^\tau$ weighted by coefficient functions $C_{\mu\nu}^{\tau,n}$ [18]. The index τ defines the twist of the operator, which corresponds to factors of $(-1/q^2)^{\tau/2-1}$ in the series. At high probing energies, leading twist operators dominate the contributions. The Mellin transformed hadronic tensor (as a function of x) appears as a series over the twist of the product of coefficient functions and operators:

$$\tilde{W}_{\mu\nu}(q^2, n) = \frac{1}{4} \sum_{\tau=2}^{\infty} C_{\mu\nu}^{\tau,n} \mathcal{O}_n^\tau \left(\frac{-1}{q^2}\right)^{\frac{\tau}{2}-1} \tag{2.12}$$

Recalling the Mellin convolution theorem, $\tilde{W}_{\mu\nu}$ appears is a Mellin transformed function that is equal to the product of functions. Interpreting $C^{\tau,n}_{\mu\nu}$ and $\mathcal{O}^{\tau}_{n}$ also as Mellin transformed functions, the leading twist contributions take the form [18]

$$\begin{aligned} \mathcal{O}^{\tau=2}_{n} &= \sum_{i=q,\bar{q},g} \tilde{f}^{H}_{i}(n) = \int \mathrm{d}y\, y^{n-1} \sum_{i=q,\bar{q},g} f^{H}_{i}(y) \\ C^{\tau=2,n}_{\mu\nu} &= \tilde{\sigma}_{\mu\nu}(n) = \int \mathrm{d}z\, z^{n-1} \sigma_{\mu\nu}(z) \end{aligned} \tag{2.13}$$

Thus, by the convolution theorem, the leading twist contribution to the hadronic tensor appears as a convolution:

$$W_{\mu\nu}(q^2, x) = \frac{1}{4} \sum_{i=q,\bar{q},g} (f^{H}_{i} \otimes \sigma_{\mu\nu})(x) + O\left(\frac{1}{q^4}\right) \tag{2.14}$$

2.1.2 Numerical Techniques

In practice, operators $\mathcal{O}(x)$ need to be convoluted with distribution functions $d(x)$. However, computing convolutions is expensive since an integral needs to be computed every time x is evaluated. To reduce the number of integrations, d is assumed to be sufficiently smooth such that it may be approximated by a collection of interpolation functions [3]. For a grid $g = \{x_0, \ldots, x_N\}$ of N interpolation points and a collection of $\{w_0, \ldots, w_N\}$ interpolation functions, the distribution may be expressed as $d(x) = \sum_{i=0}^{N} w_i(x) d_i$ where $d_i = d(x_i)$. The distribution function only needs to be evaluated at each grid point d_i. The convolution may be evaluated at a grid point by:

$$(\mathcal{O} \otimes d)(x_i) = \int_{x_i}^{1} \frac{\mathrm{d}y}{y} \mathcal{O}\left(\frac{x_i}{y}\right) d(y) = \int_{x_i}^{1} \frac{\mathrm{d}y}{y} \mathcal{O}\left(\frac{x_i}{y}\right) \sum_{j=0}^{N} w_j(y) d_j = \sum_{j=0}^{N} \mathcal{O}_{ij} d_j \tag{2.15}$$

where the operator $\mathcal{O}_{ij}$ may be precomputed on the square grid $g \times g$ as

$$\mathcal{O}_{ij} = \int_{x_i}^{1} \frac{\mathrm{d}y}{y} \mathcal{O}\left(\frac{x_i}{y}\right) w_j(y) \tag{2.16}$$

The original convolution $\mathcal{O} \otimes d$ may be interpolated in the same manner as d: $(\mathcal{O} \otimes d)(x) = \sum_{i=0}^{N} w_i(x) (\mathcal{O} \otimes d)_i$. This procedure is used by libraries such as APFEL++ [3], which will be used to compute observables dependent on convolutions of distribution functions throughout this paper. This is also allows distribution functions to be expressed as a finite number of evaluations on a grid, which separates reporting distributions from any underlying model.

2.2 Factorization Theorems

The purpose of distribution functions is to incorporate all of the non-perturbative behavior of a strong interaction. When calculating cross sections involving hadrons, the

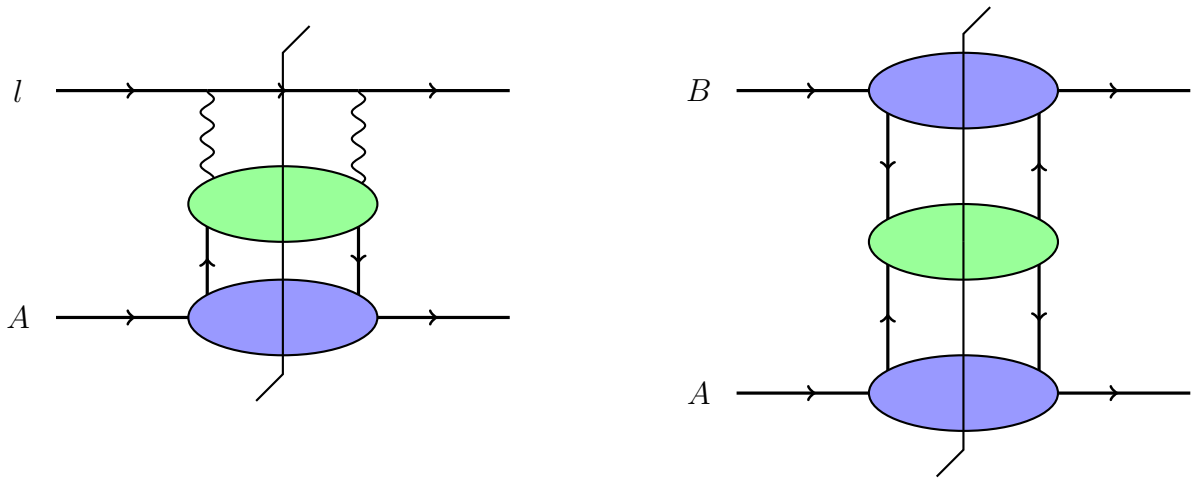

Figure 2.5: Factorizations of PDF processes. Left: factorization of DIS where a lepton l probes a hadron A. Right: factorization of hadron-hadron scattering where two hadrons A and B contribute partons to a partonic cross section.

observable may be factorized into distribution functions and a hard cross section involving just partons [6]. In the QPM, a parton is chosen from each incoming hadron with a probability given by the PDF. These partons are then involved in the partonic cross section, and hadrons are produced from the outgoing partons with a probability given by the FF. The distribution functions are combined with the partonic cross section by convolutions, and are summed over all possible choices of hadrons and partons.

Purely partonic quantities are denoted by a hat, i.e. if the hadronic cross section is σ then the partonic cross section is $\hat{\sigma}$. Cut diagrams are color coded such that the green blobs are partonic cross sections, blue blobs are PDFs, and red blobs are FFs. The colors match the defining diagrams 2.4.

The basic exclusive cross sections involving PDFs have the following factorization theorems:

- Deep Inelastic Scattering (DIS): A lepton l scatters inelastically off of a hadron A by the process $lA \to lX$
$$\sigma = \hat{\sigma} \otimes \mathrm{PDF}_A \tag{2.17}$$
- Hadron-Hadron scattering: A hadron A scatters off of a hadron B inelastically into exclusive hadrons X by the process $AB \to X$
$$\sigma = \hat{\sigma} \otimes \mathrm{PDF}_A \otimes \mathrm{PDF}_B \tag{2.18}$$

The basic cross sections involving FFs appear as semi-inclusive processes where a hadron H is observed as a final hadron. They augment the previous PDF processes, as well as including lepton-lepton annihilation:

- Semi-Inclusive Annihilation (SIA): A lepton l^- and antilepton l^+ annihilate to produce quarks through a neutral current via $l^+l^- \to HX$
$$\sigma = \hat{\sigma} \otimes \mathrm{FF}_H \tag{2.19}$$

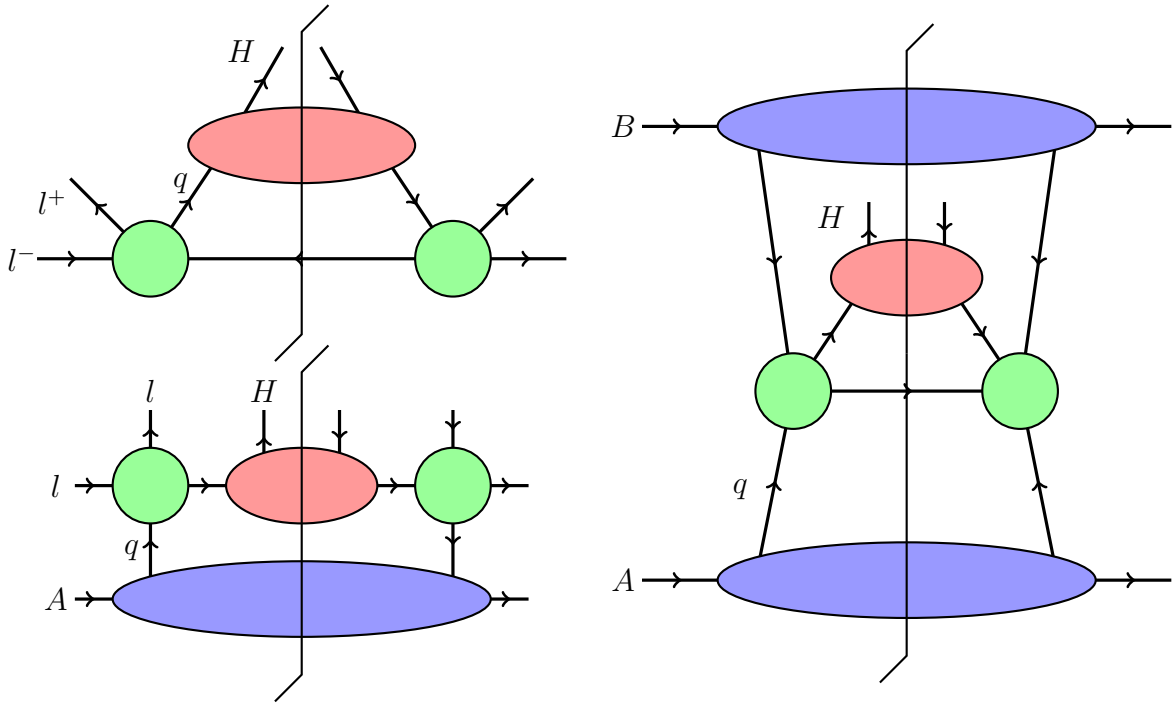

Figure 2.6: Factorization theorems of FF processes. A hadron H is observed in all out states. Upper left: factorization of SIA where a lepton-antilepton pair $l^{\pm}$ annihilate. Lower left: factorization of SIDIS where a lepton l probes a hadron A. Right: factorization of semi-inclusive hadron-hadron collisions involving hadrons A and B.

- Semi-Inclusive Deep Inelastic Scattering (SIDIS): A DIS process where H is observed in the outgoing hadrons: $lA \to lHX$

$$\sigma = \hat{\sigma} \otimes \mathrm{PDF}_A \otimes \mathrm{FF}_H \tag{2.20}$$

- Semi-Inclusive Hadron-Hadron Scattering: Hadron-hadron scattering where H is observed in the outgoing hadrons: $AB \to HX$

$$\sigma = \hat{\sigma} \otimes \mathrm{PDF}_A \otimes \mathrm{PDF}_B \otimes \mathrm{FF}_H \tag{2.21}$$

Beyond leading order, gluon exchange produces extra vertices in the propagation of partons. Soft and collinear gluon emission produces non-perturbative effects associated with the long-range physics of the strong interaction. A choice of energy scale μ_{fact} for PDFs and μ_{frag} for FFs separates low transverse momentum partons out of the partonic cross section and absorbs them into the distribution functions. The total cross section is independent of these energy scales:

$$\frac{\partial \sigma}{\partial \mu_{\mathrm{fact}}} = \frac{\partial \sigma}{\partial \mu_{\mathrm{frag}}} = 0 \tag{2.22}$$

A choice of $\mu_{\mathrm{fact}} = \mu_{\mathrm{frag}}$ is often made out of convenience, and similarly may be set equal to the re normalization scale μ of the running couplings. The specific behavior of these scales is discussed in section 2.3.

2.3 Scale Dependence and Re normalization

By incorporating QCD into the quark parton model, a partonic observable $\hat{\mathcal{O}}_0^i$ for a parton i may be calculated to higher orders in the strong coupling $\alpha_S(\mu^2)$. At NLO, qqg vertices affect both quark and gluon parton processes, and at higher orders 3- and 4-point gluon vertices contribute. Unlike electroweak loop diagrams which produce ultraviolet divergences, gluon corrections to $\hat{\mathcal{O}}_0^i$ also produce infrared divergences which arise from soft and collinear particle emission [17]. Such IR divergences are long-distance effects and so are non-perturbative. By the factorization theorems, the total observable $\mathcal{O}$ is given by convoluting the appropriate distribution functions with the partonic observable $\hat{\mathcal{O}}_0^i$ and summing over all partons. Suppose there's a single distribution type $d_0^i(x)$ to be convoluted, called the bare distribution of the parton i. Because the distribution functions already contain non-perturbative long-range effects, the IR divergences of $\hat{\mathcal{O}}_0^i$ may be transferred into d_0^i at a given scale μ^2 [18]:

$$\mathcal{O}(x,Q^2) = \sum_{i=q,\bar{q},g} \hat{\mathcal{O}}_0^i(x,Q^2) \otimes d_0^i(x) = \sum_{i=q,\bar{q},g} \hat{\mathcal{O}}^i(x,Q^2,\mu^2) \otimes d^i(x,\mu^2) \tag{2.23}$$

In the case of PDFs, μ^2 is known as the factorization scale, while for FFs, μ^2 is the fragmentation scale. Like the renormalization of masses and bare couplings, the factorization and fragmentation scales define the scale at which IR divergences are subtracted from $\hat{\mathcal{O}}^i$ and instead incorporated into d^i. The subtraction method is scheme dependent [17], e.g. the DIS scheme ensures there are no $O(\alpha_S)$ corrections to the DIS structure functions of the proton (F_2, F_3, F_L) [17]. More generally, the $\overline{\text{MS}}$ scheme only subtracts the IR divergence and a $\ln(4\pi e^{-\gamma_E})$ contribution associated with dimensional regulation calculations (here γ_E is the Euler-Mascheroni γ constant) [17]. Generally in $\overline{\text{MS}}$, partonic observables will have non-trivial $O(\alpha_S)$ contributions that will need to be calculated. The perturbative expansion of $\hat{\mathcal{O}}^i$ takes the form

$$\hat{\mathcal{O}}^i(x,Q^2,\mu^2) = A^i \sum_{n=0}^{\infty} \left(\frac{\alpha_S(\mu^2)}{2\pi}\right)^n C_n^i(x) \tag{2.24}$$

where A^i is a common prefactor, and $C_n^i(x)$ are the coefficient functions at order n for the parton i.

Analogous to renormalization, the scale dependence of the distribution functions (i.e. their "running") is given by the DGLAP equation to $O(\alpha_S)$ (i.e. the "renormalization group equations") [17]:

$$\mu^2 \frac{\partial}{\partial \mu^2} d(x,\mu^2) = \frac{\alpha_S(\mu^2)}{2\pi} P(x) \otimes d(x,\mu^2) = \frac{\alpha_S(\mu^2)}{2\pi} \int_x^1 \frac{\mathrm{d}\xi}{\xi} P\left(\frac{x}{\xi}\right) d(\xi,\mu^2) \tag{2.25}$$

As with the QPM, a more formal derivation may be done with operator product expansions of quark and gluon operators and applying the renormalization group equation to their associated anomalous dimensions [18]. This is equivalent to the DGLAP equations at $O(\alpha_S)$, but also extends them to all orders. The general DGLAP equations are:

$$\mu^2 \frac{\partial}{\partial \mu^2} \begin{pmatrix} d_{q_i} \\ d_g \end{pmatrix} = \frac{\alpha_S(\mu^2)}{2\pi} \sum_{q_j,\bar{q}_j} \begin{pmatrix} P_{q_i q_j} & P_{q_i g} \\ P_{g q_j} & P_{gg} \end{pmatrix} \otimes \begin{pmatrix} d_{q_j} \\ d_g \end{pmatrix} \tag{2.26}$$

where $d_q(x,\mu^2)$ are the quark distributions, $d_g(x,\mu^2)$ are the gluon distributions, and $P(x,\alpha_S(\mu^2)$ are the splitting functions. Each splitting function is given by the IR divergences of the partonic observable $\hat{\mathcal{O}}_0$ being shifted into the quark and gluon distributions. Since divergences occur with gluon vertices, the splitting functions also admit a perturbative series in α_S. The solutions to the evolution equations take the form of a linear evolution operator $\mathbf{\Gamma}(x;\mu^2,\mu_0^2)$ which evolves the distribution functions from a scale μ_0^2 to μ^2 [2]:

$$d^i(x,\mu^2) = \sum_{j=q,\bar{q},g} \Gamma_{ij}(x;\mu^2,\mu_0^2) \otimes d^j(x,\mu_0^2) \tag{2.27}$$

As with previous numerical techniques to calculate observables dependent on distribution functions, the evolution operator may be calculated on an x grid independently of d for a fixed energy scale μ_0^2 to μ^2.

Corrections to a partonic observable $\hat{\mathcal{O}}^i$ may also include electroweak interactions. While the EW coupling α is significantly smaller than the strong coupling α_S, at high energies the exchange of soft and collinear massive gauge bosons may produce logarithms of the order $\ln^2(Q^2/M_W^2)$ [19]. The same analysis of splitting functions in QCD may be applied to EW interactions, and thus EW splitting functions may be calculated for the quarks, leptons, photons, and weak massive bosons [2]. A corresponding evolution equation involving EW splitting functions applies to the distribution functions now at a EW scale ν^2:

$$\nu^2 \frac{\partial}{\partial \nu^2} \begin{pmatrix} d_{q_i} \\ d_g \end{pmatrix} = \frac{\alpha(\nu^2)}{2\pi} \sum_{q_j,\bar{q}_j} \begin{pmatrix} P^{EW}_{q_i q_j} & P^{EW}_{q_i g} \\ P^{EW}_{g q_j} & P^{EW}_{gg} \end{pmatrix} \otimes \begin{pmatrix} d_{q_j} \\ d_g \end{pmatrix} \tag{2.28}$$

The solution to the EW evolution equations is likewise a linear evolution operator $\mathbf{\Gamma}^{\mathrm{EW}}(x;\nu^2,\nu_0^2)$ satisfying

$$d^i(x,\nu^2) = \sum_{j=q,\bar{q},g} \Gamma^{\mathrm{EW}}_{ij}(x;\nu^2,\nu_0^2) \otimes d^j(x,\nu_0^2) \tag{2.29}$$

If the EW and QCD evolution equations are solved simultaneously, the evolution operator takes the form $\mathbf{\Gamma}^{\mathrm{QCD}\otimes\mathrm{EW}}$. However when solved independently, the order of individual evolution functions produces $O(\alpha\alpha_S)$ errors [2]. Taking the average of QCD then EW evolution and EW then QCD evolution cancels this linear order error, and thus has accuracy up to $O(\alpha^2)$.

At higher orders, the evolutions have non-trivial mixing between different flavors which makes solving the integrodifferential equations difficult. By using $SU(n_f)$ flavor symmetry, the flavor structure of the splitting functions may be expressed in terms of singlet S and non-singlet V components [2]:

$$\begin{aligned} P_{q_i q_j} &= \delta_{ij} P^V_{qq} + P^S_{qq} \\ P_{q_i \bar{q}_j} &= \delta_{ik} P^V_{q\bar{q}} + P^S_{q\bar{q}} \\ P^{\pm} &= P^V_{qq} \pm P^V_{q\bar{q}} \\ P^S_{qq} &= P^S_{q\bar{q}} \end{aligned} \tag{2.30}$$

The flavor basis may then be partitioned into a non-singlet basis:

$$
\begin{array}{ll}
V_i = q_i^- & T_{15} = u^+ + d^+ + s^+ - 3c^+ \\
T_3 = u^+ - d^+ & T_{24} = u^+ + d^+ + s^+ + c^+ - 4b^+ \\
T_8 = u^+ + d^+ - 2s^+ & T_{35} = u^+ + d^+ + s^+ + c^+ + b^+ - 5t^+
\end{array}
$$

where $q^\pm = q \pm \overline{q}$. This non-singlet basis is entirely decoupled between each of its basis functions, as well as from the singlet basis:

$$
\begin{pmatrix} \Sigma = u^+ + d^+ + s^+ + c^+ + b^+ + t^+ \\ g \end{pmatrix} \tag{2.31}
$$

which is a 2D coupled system. For EW, a basis consisting of a (different) singlet $\boldsymbol{q}^{\mathrm{SG}}$ and non-singlet q_i^{NS} set decouples the system:

$$
\begin{aligned}
\boldsymbol{q}^{\mathrm{SG}} &= \begin{pmatrix} \gamma \\ \Sigma = u^+ + d^+ + s^+ + c^+ + b^+ + t^+ \\ D_{\Delta\Sigma} = u^+ + c^+ + t^+ - d^+ - s^+ - b^+ \end{pmatrix} \\
q^{\mathrm{NS}} &= \left\{ D_{uc}, D_{ds}, D_{sb}, D_{ct}, u^-, d^-, s^-, c^-, b^-, t^- \right\}
\end{aligned} \tag{2.32}
$$

where $D_{ab} = a^+ - b^+$. The singlet functions form a 3x3 coupled system, while the 10 non-singlet functions are decoupled. Given a QCD basis, the transformation to the EW basis is a rotation by an invertible linear transformation. Since the evolution equations are linear, the change of basis transformation may be absorbed into the evolution operator. Thus the evolution operators may be discussed with respect to the original flavor basis.

2.3.1 Time-like Splitting Functions

While both PDFs and FFs follow the same general analysis outlined above, the functional forms of splitting functions for FFs (time-like) are different than the splitting functions of PDFs (space-like). Firstly, for a space-like splitting function P_{ij}, the corresponding time-like splitting function is denoted P_{ji}. Thus the singlet time-like evolution equation is often written

$$
\mu^2 \frac{\partial}{\partial \mu^2} \begin{pmatrix} D_\Sigma \\ D_g \end{pmatrix} = \frac{\alpha_S(\mu^2)}{2\pi} \begin{pmatrix} P_{qq} & 2n_f P_{gq} \\ P_{qg} & P_{gg} \end{pmatrix} \otimes \begin{pmatrix} D_\Sigma \\ D_g \end{pmatrix} \tag{2.33}
$$

At leading order, the time-like and space-like splitting functions are indeed equal, but including NLO terms introduces singularities as $x \to 0$ in the gluon splitting functions that are not present in the space-like splitting functions [17]. These singularities are due to soft gluon emission, but may be resummed in power series of α_S to provide a coherent final state where soft gluons are suppressed [17].

2.4 Semi-Inclusive Annihilation

Semi-inclusive lepton-antilepton annihilation provides the most direct access to fragmentation functions. Since there are no initial state hadrons, there are no PDFs, so all of the non-perturbative strong effects are in the FF. The process is normally expressed as

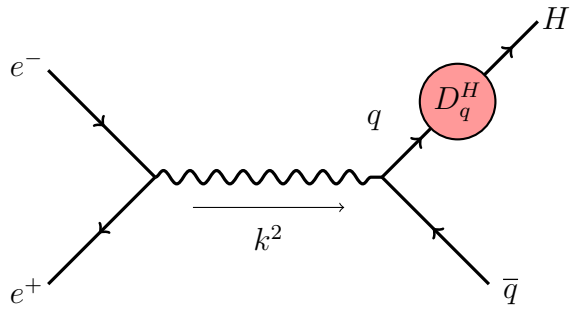

Figure 2.7: Tree level process of SIA involving electron-positron annihilation

electron-position annihilation $e^-e^+ \to X$, and the total exclusive process as a ratio of hadron production vs muon production for a center of mass energy $\sqrt{s}$ [17]:

$$\begin{aligned} R &= \frac{\sigma(e^+e^- \to X)}{\sigma(e^+e^- \to \mu^+\mu^-)} = 3\sum_q Q_q^2 \qquad s \ll M_Z^2 \\ R_Z &= \frac{\sigma(Z \to X)}{\sigma(Z \to \mu^+\mu^-)} = \frac{3\sum_q (A_q^2 + V_q^2)}{A_\mu^2 + V_\mu^2} \qquad s \approx M_Z^2 \end{aligned} \tag{2.34}$$

where Q, A, and V are the photon, axial, and vector couplings respectively. The color structure is entirely contained as the prefactor of 3. While (2.34) is to leading order in α_S, its expansion occurs as a multiplicative correction:

$$R = 3\sum_q Q_q^2 \to 3\sum_q Q_q^2 \left(1 + \frac{\alpha_S}{\pi} + O(\alpha_S^2)\right) = K_{\text{QCD}} R \tag{2.35}$$

and the same for R_Z. Higher order corrections to K_{QCD} have been calculated, but introduce a dependence on the number of active quark flavors n_f.

The semi-inclusive process is identical, with the addition of identifying an outgoing hadron H. From the definition of FFs, the normalized differential cross section for SIA may be

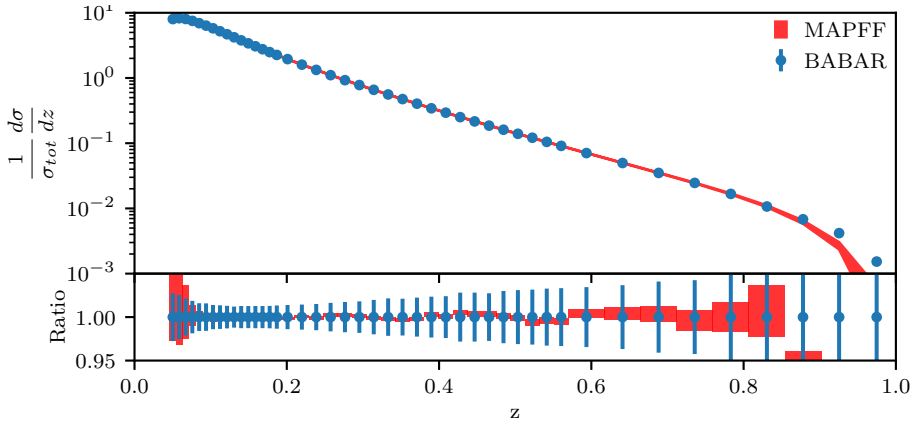

Figure 2.8: BABAR data (blue) plotted against predictions made by MAPFF (red). The bands of the MAPFF fit corresponds to the variance of the observable.

predicted by combining (2.4) and (2.5):

$$\frac{1}{\sigma_{\text{tot}}}\frac{\mathrm{d}\sigma}{\mathrm{d}z} = \sum_{i=q,\bar{q},g} C_i(s;z,\alpha_S) \otimes D_i^H(z,s) \tag{2.36}$$

where $\sigma_{\text{tot}} = R\sigma(e^+e^- \to \mu^+\mu^-)$ below the Z threshold, or the corresponding Z variables near the Z pole.

A sample comparison of SIA differential cross sections from BABAR [20] is plotted in 2.8 against predictions made by `MAPFF1.0` for charged pion production [5]. The calculations were performed using APFEL [2], which provides built-in methods of computing SIA cross sections and structure functions. Note that the center of mass energy $\sqrt{s} = 10.54$ GeV is below $M_Z \approx 91.2$ GeV, so all calculations were done in QED. Because of the simplicity of the process, the terms may be calculated to perturbative orders which give highly constrained predictions. For further analysis and fitting of SIA data, refer to [4].

2.5 Semi-Inclusive Deep Inelastic Scattering

Deep inelastic scattering provides the simplest process that gives rise to PDFs. Unpolarized observables from DIS are often reported in terms of structure functions F_2 and F_L. For a lepton scattering off of a proton via a photon with virtuality $-q^2 = Q^2 > 0$ shown in 2.9, the differential cross section is given as [18]:

$$\frac{\mathrm{d}^2\sigma}{\mathrm{d}x\,\mathrm{d}Q^2} = \frac{4\pi\alpha^2}{xQ^4}\left(\left[1+(1-y)^2\right]F_2(x,Q^2) - y^2F_L(x,Q^2)\right) \tag{2.37}$$

where $x = Q^2/2p\cdot q$ is the Bjorken x-variable and $y = p\cdot q/k\cdot p$. The structure functions are calculated as convolutions of coefficient functions with PDFs:

$$F_j(x,\mu^2) = \sum_{i=q,\bar{q},g} C_j(x,\mu^2)\otimes f_i(x,\mu^2) \qquad\qquad j = 2, L \tag{2.38}$$

Coefficient functions are readily available to NLO [21].

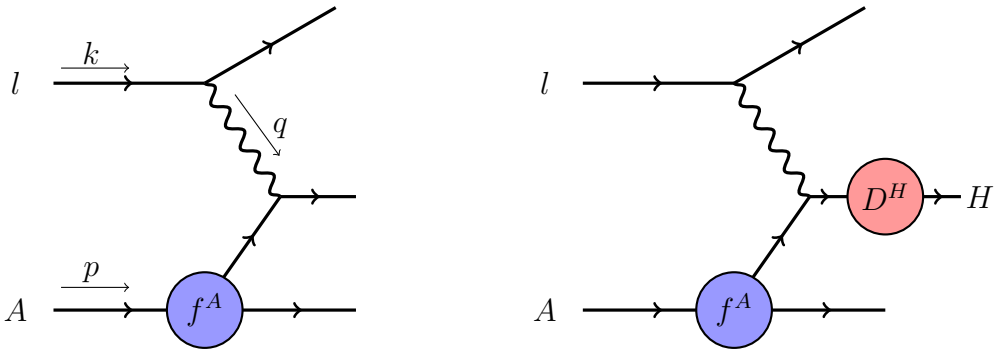

Figure 2.9: Left: Tree level diagram of DIS. Right: Tree level diagram of SIDIS.

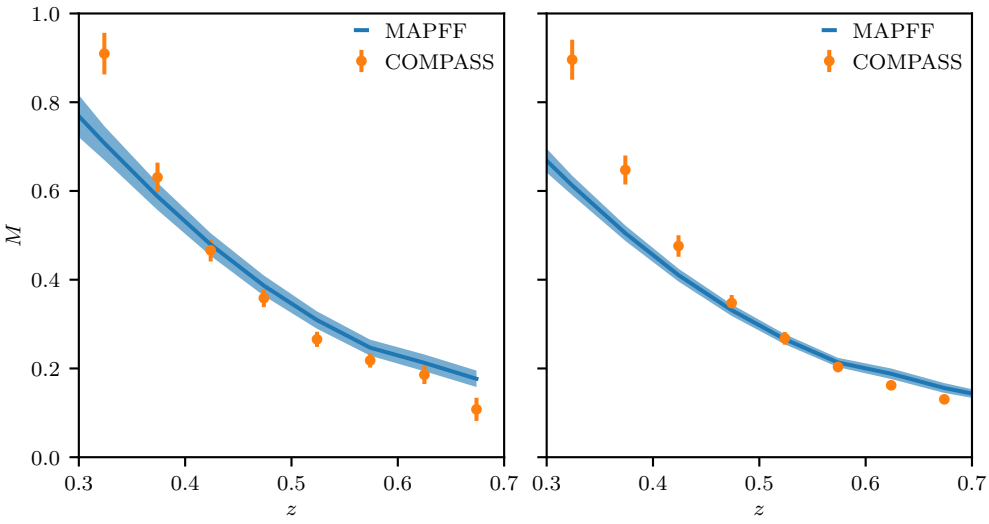

Figure 2.10: Comparison of COMPASS and MAPFF integrated multiplicities from equation (2.40). Both plots are integrated over a y bin of 0.3 to 0.5 and: Left: x bin of 0.004 to 0.01. Right: x bin of 0.01 to 0.02. Plots serve as a check of `MAPFF` results but do not include all corrections from [5].

The SIDIS differential cross section takes a similar form where the structure functions have been convoluted with the fragmentation functions $D_q^H(z)$:

$$\begin{aligned} \frac{\mathrm{d}^3\sigma}{\mathrm{d}z\,\mathrm{d}Q^2\,\mathrm{d}z} &= \frac{4\pi\alpha^2}{xQ^4}\left(\left[1+(1-y)^2\right]F_2(x,z,Q^2) - y^2 F_L(x,z,Q^2)\right) \\ F_k(x,z,\mu^2) &= \sum_{i,j=q,\bar{q},g} C_{k,ij}(x,z,\mu^2)\otimes f_i(x,\mu^2)\otimes D_j^H(z,\mu^2), \qquad k=2,L \end{aligned} \tag{2.39}$$

The observables are now dependent upon two non-perturbative distribution functions: $f_i(x,\mu^2)$ and $D_i^H(z,\mu^2)$. Because of the large amount of data used to fit PDFs independently of FFs, the PDFs may be treated as inputs of the theory rather than quantities to be fitted. The amount of data points lost by not fitting against SIDIS data concurrently with FFs may be considered negligible. SIDIS data is often delivered as a multiplicity M, which is the ratio of integrated cross sections:

$$M = \left[\int_{Q_{\min}}^{Q_{\max}}\mathrm{d}Q\int_{x_{\min}}^{x_{\max}}\mathrm{d}x\int_{z_{\min}}^{z_{\max}}\mathrm{d}z\frac{\mathrm{d}^3\sigma}{\mathrm{d}x\,\mathrm{d}Q\,\mathrm{d}z}\right] \Big/ \left[\int_{Q_{\min}}^{Q_{\max}}\mathrm{d}Q\int_{x_{\min}}^{x_{\max}}\mathrm{d}x\frac{d^2\sigma}{\mathrm{d}x\,\mathrm{d}Q}\right] \tag{2.40}$$

A sample comparison of multiplicities from COMPASS data [22] is plotted in figure 2.10 against predictions made by `MAPFF1.0` for charged pion production. The calculations were performed using `APFEL++` [3], which extends the `APFEL` library to perform numeric convolutions of arbitrary operators and distributions. The calculations closely follow the techniques in [5]. The PDF was similarly chosen to be `NNPDF31_as_nlo_pch_as_0118` [11].

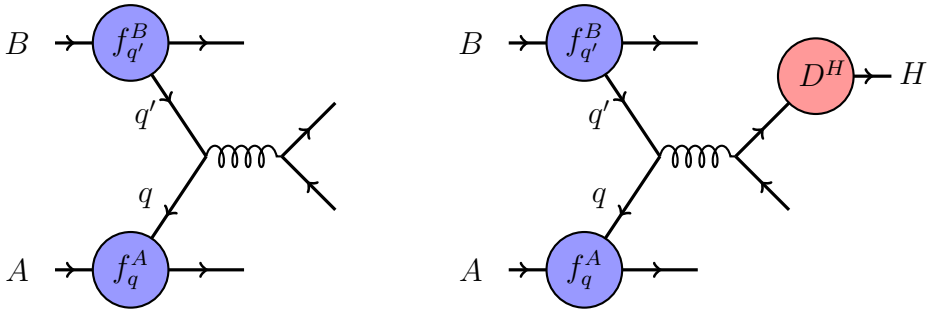

Figure 2.11: Tree level diagram of hadron-hadron scattering (Left: exclusive, Right: semi-inclusive). Both diagrams show only one of the leading order partonic processes.

2.6 Semi-Inclusive Hadron-Hadron Scattering

Scattering of hadrons introduces two non-perturbative distributions (PDFs) in the initial hadrons, which makes fitting increasingly sensitive to measurement uncertainties. The simplest expression for the differential cross section in figure 2.11 is just the factorization theorem for $AB \to X$:

$$\frac{\mathrm{d}^3\sigma}{\mathrm{d}p^3} = \sum_{i,j=q,\bar{q},g} \frac{\mathrm{d}^3\hat{\sigma}_{ij}}{\mathrm{d}p^3} \otimes f_i^A \otimes f_j^B \tag{2.41}$$

The leading order partonic processes are shown in figure 2.12. Of these, the gluon-gluon processes dominate due to the larger number of color combinations available for sampling [9]. Therefore hadron-hadron scattering provides a sensitive measurement of the gluon distributions.

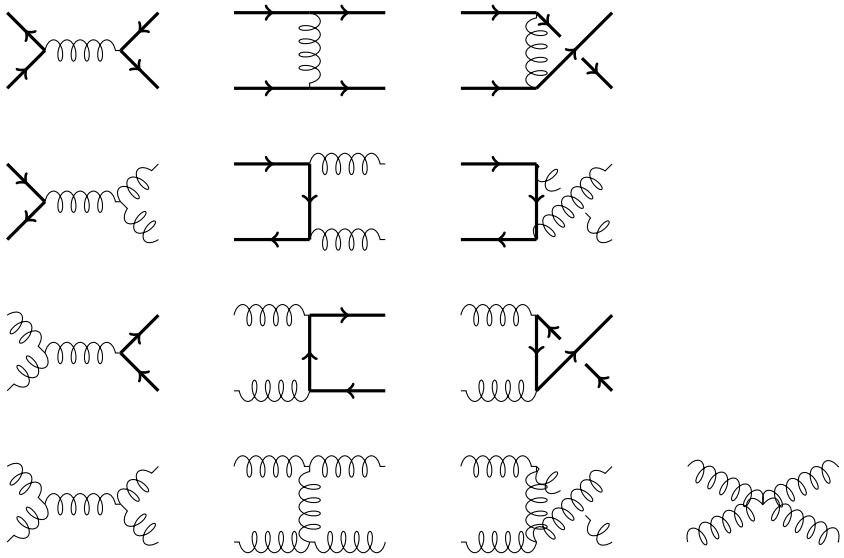

Figure 2.12: Leading order α_S partonic processes that contribute to hadron-hadron scattering. Rows are organized by the number of quark-antiquark pairs in the in and out states. The final row also includes the gluon 4-point cross section.

The semi-inclusive hadron-hadron scattering also follows from factorization, and provides

the same sensitivity to gluon distributions:

$$\frac{\mathrm{d}^3\sigma}{\mathrm{d}p^3} = \sum_{i,j,k=q,\bar{q},g} \frac{\mathrm{d}^3\hat{\sigma}_{ij\to kX}}{\mathrm{d}p^3} \otimes f_i^A \otimes f_j^B \otimes D_k^H \tag{2.42}$$

The focus of this paper will be on the analysis of results computed by equation (2.42). Recent semi-inclusive hadron-hadron scattering measurements are mostly proton-proton collisions, so the PDFs will be assumed to be PDFs of the proton. Numerical techniques discussed in section 2.1.2 allow the differential cross section to be computed directly, so further analysis is superfluous here. Those interested may consult [17] for photon production and the photon fragmentation function, or [8] for studying small-x behavior.

Chapter 3

Experimental

This chapter explains the parameters, observables, and uncertainties of experimental data sets and how to calculate them from provided data. The experimental data sets used in reweighting are explained here, with details regarding particular experiments enumerated.

3.1 Parameters

Consider a particle beam of center of mass energy $\sqrt{s}$ directed along the z-axis. Particles produced by scattering events are observed in a finite cylinder about the beam axis. An out-state particle momentum has cylindrical symmetry, and thus its transverse magnitude may be determined by a radial component p_T and angle ϕ. The momentum along the beam axis is commonly reported as the rapidity y:

$$y = \ln \sqrt{\frac{E + p_z c}{E - p_z c}} = \frac{1}{2} \ln \frac{p^+}{p^-} \tag{3.1}$$

where $p^\pm$ are the light-cone coordinates [17]. At high energy, this is equivalent to the pseudorapidity η:

$$\eta = -\ln \tan \frac{\theta}{2} \tag{3.2}$$

where θ is the angle between the outgoing particle and the beam axis. The original momentum may be derived from p_T, ϕ, and y via [17]:

$$p^\mu = (m_T \cosh y, p_T \sin \phi, p_T \cos \phi, m_T \sinh y) \tag{3.3}$$

where $m_T = \sqrt{m^2 + p_T^2}$. The corresponding change of variable volume is given by:

$$\mathrm{d}^3 p = 2\pi E \ \mathrm{d}y \, \mathrm{d}p_T^2 \tag{3.4}$$

3.2 Observables

As derived in the theory section, the semi-inclusive proton-proton scattering into a pion is given by

$$\begin{aligned}\sigma(P_1 + P_2 \to P_3 + X) = \sum_{ijk} \int_0^1 \mathrm{d}x_1\, \mathrm{d}x_2 \frac{\mathrm{d}z}{z} f_i(x_1, \mu^2_{\text{fact}}) f_j(x_2, \mu^2_{\text{fact}}) D^\pi_k(z, \mu^2_{\text{frag}}) \\ \times \sigma(\hat{p}_1 + \hat{p}_2 \to \hat{p}_3, \mu^2_{\text{ren}}, \mu^2_{\text{fact}}, \mu^2_{\text{frag}})\end{aligned} \tag{3.5}$$

for the proton PDFs $f_i(x, \mu^2)$ and the pion's fragmentation functions $D^\pi_i(z, \mu^2)$. For a 2-2 scattering partonic process, the differential cross section is defined as

$$E_3 E_4 \frac{\mathrm{d}^6 \hat{\sigma}}{\mathrm{d}p_3^3\, \mathrm{d}p_4^3} = \frac{1}{2\hat{s}} \frac{1}{16\pi^2} \overline{\sum} |\mathcal{M}|^2 \delta^4(p_1 + p_2 - p_3 - p_4) \tag{3.6}$$

where $\overline{\sum}$ is an appropriately averaged sum over spin states, and $\mathcal{M}(ij \to kl)$ is the 2-2 scattering process

$$\text{parton}_i(p_1) + \text{parton}_j(p_2) \to \text{parton}_k(p_3) + \text{parton}_l(p_4) \tag{3.7}$$

Parton 4 which doesn't fragment into the pion may have its momentum integrated over to give

$$E_3 \frac{\mathrm{d}^3 \hat{\sigma}}{\mathrm{d}p_3^3} = \frac{1}{2\hat{s}} \frac{1}{8\pi^2} \overline{\sum} |\mathcal{M}|^2 \delta(\hat{s} + \hat{t} + \hat{u}) \tag{3.8}$$

for the standard partonic Mandelstam variables $\hat{s}, \hat{t}, \hat{u}$. The total cross section is given by directly substituting (3.8) into (3.5):

$$\begin{aligned}E_3 \frac{\mathrm{d}^3 \sigma}{\mathrm{d}p_3^3} = \frac{1}{16\pi^2 s} \sum_{ijkl} \int_0^1 \frac{\mathrm{d}x_1}{x_1} \frac{\mathrm{d}x_2}{x_2} \frac{\mathrm{d}z}{z} f_i(x_1, \mu^2_{\text{fact}}) f_j(x_2, \mu^2_{\text{fact}}) D^\pi_k(z, \mu^2_{\text{frag}}) \\ \times \overline{\sum} |\mathcal{M}(ij \to kl)|^2 \delta(\hat{s} + \hat{t} + \hat{u})\end{aligned} \tag{3.9}$$

The cross section is also given in terms of an equivalent parameterization of y and p_T:

$$E_3 \frac{\mathrm{d}^3 \sigma}{\mathrm{d}p_3^3} = \frac{\mathrm{d}^3 \sigma}{\mathrm{d}y\, \mathrm{d}p_T^2} = \frac{1}{2\pi p_T} \frac{\mathrm{d}^2 \sigma}{\mathrm{d}y\, \mathrm{d}p_T} \tag{3.10}$$

3.3 Experimental Data Sets

The analysis of MAPFF is based upon semi-inclusive proton-proton collisions given by the following process:

$$p + p \to \pi + X \tag{3.11}$$

where π may be charged, neutral, or the sum of positive and negatively charged pions. The data is listed as differential cross sections (3.10) dependent on the center of mass energy $\sqrt{s}$, the inclusive pion's rapidity y, and its transverse momentum p_T. Data is

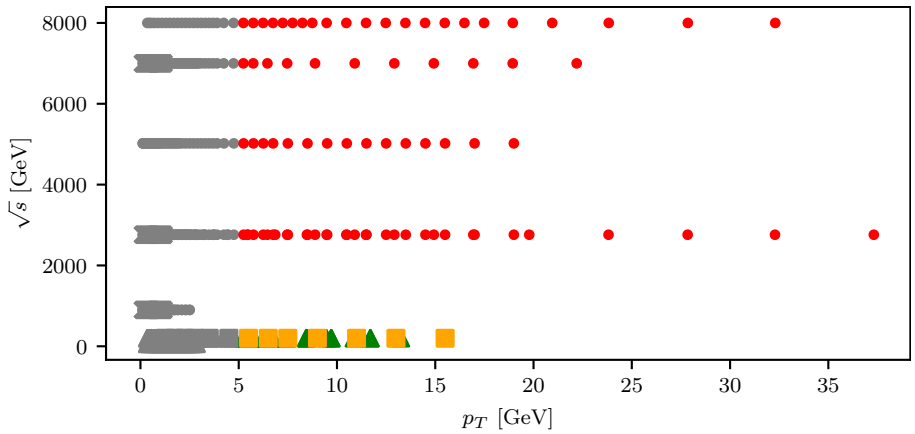

Figure 3.1: Kinematic coverage of the experimental data. Grey: Data below momentum cutoff. Red dots: ALICE experiments. X's: CMS experiments (all cut). Green triangles: PHENIX experiments. Orange squares: STAR experiments.

Name	Ref.	Pion	Observable	N_{dat}	$\sqrt{s}$ [GeV]	p_T-range	y
ALICE	[23]	$\pi^{\pm}$	$\frac{1}{N_{\text{ev}}}\frac{\mathrm{d}^2\sigma}{\mathrm{d}y\,\mathrm{d}p_T}$	33	900	$0.1-2.6$	$\lvert y\rvert<0.5$
	[24]	π^0	$E\frac{\mathrm{d}^3\sigma}{\mathrm{d}p^3}$	18	2760	$0.4-10.0$	$y=0$
	[25]	π^0	$E\frac{\mathrm{d}^3\sigma}{\mathrm{d}p^3}$	30	2760	$0.4-40.0$	$\lvert y\rvert<0.6$
	[26]	$\pi^+ + \pi^-$	$\frac{1}{2\pi p_T N_{\text{ev}}}\frac{\mathrm{d}^2\sigma}{\mathrm{d}y\,\mathrm{d}p_T}$	63	2760	$0.1-20.0$	$\lvert y\rvert<0.8$
	[27]	$\pi^+ + \pi^-$	$\frac{1}{N_{\text{ev}}}\frac{\mathrm{d}^2\sigma}{\mathrm{d}y\,\mathrm{d}p_T}$	58	5020	$0.1-20.0$	$\lvert y\rvert<0.5$
	[28]	π^0	$E\frac{\mathrm{d}^3\sigma}{\mathrm{d}p^3}$	33	7000	$0.3-25.0$	$y=0$
	[29]	$\pi^+ + \pi^-$	$\frac{1}{N}\frac{d^2\sigma}{dydp_T}$	41	7000	$0.1-3.0$	$\lvert y\rvert<0.5$
	[30]	π^0	$E\frac{\mathrm{d}^3\sigma}{\mathrm{d}p^3}$	44	8000	$0.3-35.0$	$y=0$
CMS	[35]	$\pi^{\pm}$	$\frac{1}{N_{\text{ev}}}\frac{\mathrm{d}^2\sigma}{\mathrm{d}y\,\mathrm{d}p_T}$	22	900	$0.125-1.175$	$\lvert y\rvert<1.0$
	[35]	$\pi^{\pm}$	$\frac{1}{N_{\text{ev}}}\frac{\mathrm{d}^2\sigma}{\mathrm{d}y\,\mathrm{d}p_T}$	22	2760	$0.125-1.175$	$\lvert y\rvert<1.0$
	[35]	$\pi^{\pm}$	$\frac{1}{N_{\text{ev}}}\frac{\mathrm{d}^2\sigma}{\mathrm{d}y\,\mathrm{d}p_T}$	22	7000	$0.125-1.175$	$\lvert y\rvert<1.0$
PHENIX	[31]	$\pi^{\pm}$	$E\frac{\mathrm{d}^3\sigma}{\mathrm{d}p^3}$	26	62	$0.3-2.9$	$\lvert y\rvert<0.6$
	[31]	$\pi^{\pm}$	$E\frac{\mathrm{d}^3\sigma}{\mathrm{d}p^3}$	27	200	$0.3-3.0$	$\lvert y\rvert<0.6$
	[32]	$\pi^{\pm}$	$E\frac{\mathrm{d}^3\sigma}{\mathrm{d}p^3}$	6	200	$5.0-13.0$	$\lvert y\rvert<0.35$
	[33]	π^0	$E\frac{\mathrm{d}^3\sigma}{\mathrm{d}p^3}$	17	200	$1.0-15.0$	$\lvert y\rvert<0.35$
	[34]	π^0	$E\frac{\mathrm{d}^3\sigma}{\mathrm{d}p^3}$	28	510	$1.0-30.0$	$\lvert y\rvert<0.35$
STAR	[36]	π^0	$E\frac{\mathrm{d}^3\sigma}{\mathrm{d}p^3}$	13	200	$1.0-17.0$	$0<y<1.0$
	[37]	π^0	$E\frac{\mathrm{d}^3\sigma}{\mathrm{d}p^3}$	13	200	$1.0-17.0$	$0.8<y<2.0$

Table 3.1: Details of unfiltered experimental observables and coverage.

Name	Ref.	Pion	Observable	N_{dat}	$\sqrt{s}$ [GeV]	p_T-range	y
ALICE	[24]	π^0	$E\frac{\mathrm{d}^3\sigma}{\mathrm{d}p^3}$	4	2760	$5.0 - 10.0$	$y = 0$
	[25]	π^0	$E\frac{\mathrm{d}^3\sigma}{\mathrm{d}p^3}$	15	2760	$5.0 - 40.0$	$\lvert y\rvert < 0.6$
	[28]	π^0	$E\frac{\mathrm{d}^3\sigma}{\mathrm{d}p^3}$	11	7000	$5.0 - 25.0$	$y = 0$
	[30]	π^0	$E\frac{\mathrm{d}^3\sigma}{\mathrm{d}p^3}$	22	8000	$5.0 - 35.0$	$y = 0$
PHENIX	[32]	$\pi^{\pm}$	$E\frac{\mathrm{d}^3\sigma}{\mathrm{d}p^3}$	6	200	$5.0 - 13.0$	$\lvert y\rvert < 0.35$
	[33]	π^0	$E\frac{\mathrm{d}^3\sigma}{\mathrm{d}p^3}$	9	200	$5.0 - 15.0$	$\lvert y\rvert < 0.35$
	[34]	π^0	$E\frac{\mathrm{d}^3\sigma}{\mathrm{d}p^3}$	20	510	$5.0 - 30.0$	$\lvert y\rvert < 0.35$
STAR	[36]	π^0	$E\frac{\mathrm{d}^3\sigma}{\mathrm{d}p^3}$	7	200	$5.0 - 17.0$	$0 < y < 1.0$
	[37]	π^0	$E\frac{\mathrm{d}^3\sigma}{\mathrm{d}p^3}$	7	200	$5.0 - 17.0$	$0.8 < y < 2.0$
Total				101			

Table 3.2: Details of filtered experimental observables and coverage.

provided by experiments at ALICE [23] [24] [25] [26] [27] [28] [29] [30], PHENIX [31] [32] [33] [34], CMS [35], and STAR [36] [37].

The kinematic coverage of the experiments is shown in figure 3.1. Details of each experiment are listed in table 3.1. For experiments in which the rapidity was binned in a specific interval, the rapidity was integrated over said interval. Otherwise the cross sections were directly evaluated at the rapidity point.

Fiducial cuts were made over low p_T, specifically $p_T^{\min} = 5$ GeV. This is due to the limited predictive ability that fragmentation functions have at low transverse momentum [16]. The remaining data after filtering low p_T is given in table 3.2. For an analysis of the dependence of the fit against the choice of momentum cut, see the 5.4.

While many ALICE experiments provide heavy ion (i.e. lead) collision data [24] [26] [27], only proton-proton collision data was used in this fitting. This avoids nuclear effects in favor of the simpler proton structure, but heavy ion collisions may be included in the future if so desired. Furthermore, the ALICE experiments which only provided a normalized observable were too expensive to fit against, and were not included. ALICE experiments provided the largest center of mass energy sampling from 2760 GeV to 8000 GeV, as well as the majority of data points. These experiments also provided the most constrained uncertainty measurements and smallest normalization uncertainties. Thus the ALICE data has the greatest influence on the reweighting.

The CMS data only sampled a p_T range up to 0.175 GeV, which was far below all considered p_T cutoffs, and thus was not included.

PHENIX measurements provided the majority of low energy $\sqrt{s} = 200, 510$ GeV measurements. While charged pion data remained post-filtering, the small quantity of data points was deemed to be too small to fit the charged fragmentation functions individually, and instead the data was summed and treated as a summed pion data set.

STAR data provided the same $\sqrt{s} = 200$ GeV coverage as PHENIX, however it only recorded forward rapidities. As with ALICE, STAR also measured heavy ion (deuterium-gold collisions), but these were omitted from this analysis.

3.3.1 Covariance Matrix

In order to calculate the best-fit measurement χ^2 [38], the covariance matrix for each experiment must be calculated. While some experiments may provide their own pre-calculated covariance matrix, the filtered experiments did not, and instead provided a set of uncertainties. Call an experiment's set of N_{dat} data points $\{y_i\}$. These uncertainties were categorized as:

- Uncorrelated statistical uncertainties $\{\sigma_{\text{stat},i}\}$.
- Uncorrelated systematic uncertainties $\{\sigma_{\text{sys},i}\}$. Reported multiplicative systematic uncertainties $\{\sigma'_{\text{sys},i}\}$ are multiplied by the corresponding data point: $\sigma_{\text{sys},i} = \sigma'_{\text{sys},i} y_i$.
- Fully correlated normalization uncertainties $\{\sigma_{\text{norm},i}\}$. In the case of an overall normalization uncertainty N, then $\sigma_{\text{norm},i} = N y_i$.

where $i = 1, \ldots, N_{\text{dat}}$. The covariance matrix represents the uncorrelated and correlated uncertainties as quadratic sums:

$$\text{cov}_{ij} = (\sigma^2_{\text{stat},i} + \sigma^2_{\text{sys},i})\delta_{ij} + \sigma_{\text{norm},i}\sigma_{\text{norm},j} \qquad i, j = 1, \ldots, N_{\text{dat}} \tag{3.12}$$

Assuming all uncertainties are positive, the covariance matrix is real, positive-definite, and orthogonal. Therefore the covariance matrix is invertible, and its inverse may be quickly calculated via Cholesky decomposition for large data sets.

Chapter 4

Methodology

This chapter details the specific calculations used to fit and reweight distribution functions. A definition of fitting a continuous function from data is provided, and common methods of parameterizing distribution fits are discussed. The core of the paper, reweighting, is discussed in detail. Information loss and unweighting are outlined at the end. Extra methods of fit validation are reserved for the appendices.

Throughout this section the following notation will be used: Bold-face expresses a vector / array of data. Let D be a fixed data set

- N_D - the number of measurements in D
- $\{\boldsymbol{x}_i\}$ for $i = 1, \ldots, N_D$ - the dependent variables of D
- $\{y_i\}$ for $i = 1, \ldots, N_D$ - the independent variables of D
- cov_{ij} - the covariance matrix of D

4.1 Fitting

In the most general sense, distribution functions $d(x, Q^2)$ are elements of a continuous functional space with $x \in [0, 1]$ and $Q^2 > 0$. Fitting distributions against data sets presents a problem, since experimental data forms a finite set while distributions live in an infinite-dimensional space. Furthermore, the data sets carry an associated uncertainty which needs to be propagated to the fitted distribution. These errors may be quantified by defining a probability measure $\mathcal{P}(d)$ on the space of distributions such that the expectation value of any observable $\mathcal{O}$ matches the average over distributions weighted by this probability measure [14]:

$$\langle \mathcal{O}[d] \rangle = \int \mathrm{d}d \, \mathcal{O}[d] \mathcal{P}(d) \tag{4.1}$$

There are many choices of probability measures which are equivalent in reproducing the expectation value of an observable [14]. Fixing a choice of a probability measure, the

quality of a fit may be determined by a best fit measurement $E[d]$, which quantifies the deviation of theoretical predictions from a given data set. A standard choice of best fit is the χ^2 function:

$$\chi^2[d] = \sum_{i,j=1}^{N_D} (y_i - \mathcal{O}[d](\boldsymbol{x}_i))(\text{cov}^{-1})_{ij}(y_j - \mathcal{O}[d](\boldsymbol{x}_j)) \tag{4.2}$$

Minimizing $E[d]$ minimizes the distance between a measured point y_i and its corresponding prediction $\mathcal{O}[d](\boldsymbol{x}_i)$. The associated uncertainty, and thus probability measure, is given by expanding in a confidence region about the minimum.

4.1.1 Hessian Method

Analytically, this may be performed by the Hessian method [13], which is a quadratic expansion about the minimum of $E[d]$. Recall that a linear expansion about an extremum is zero, and thus provides no information, so a quadratic expansion is the lowest non-trivial expansion. A quadratic expansion is accurate only within some neighborhood of the minimum defined by a tolerance parameter T:

$$\Delta E[d] \leq T^2 \tag{4.3}$$

Let $d(x, Q^2; \boldsymbol{a})$ be a model of a distribution function $d(x, Q^2)$ dependent on a set of parameters $\boldsymbol{a} = \{a_k\}$ for $k = 1, \ldots, N_p$, where N_p is the number of parameters. To calculate the Hessian uncertainty of the parameters, first determine a set $\boldsymbol{a}_0$ which minimizes $E[d]$:

$$\left.\frac{\partial E[d(x, Q^2; \boldsymbol{a})]}{\partial \boldsymbol{a}}\right|_{\boldsymbol{a}=\boldsymbol{a}_0} = 0 \tag{4.4}$$

Expanding the best fit to second order in the parameters gives [13]

$$\Delta E[d(x, Q^2; \boldsymbol{a})] = E[d(x, Q^2; \boldsymbol{a})] - E[d(x, Q^2; \boldsymbol{a}_0)] \approx \sum_{i,j=1}^{N_p} H_{ij}(a_i - a_i^0)(a_j - a_j^0) \tag{4.5}$$

where H is the Hessian matrix. Since Hessian matrices are real and symmetric, by the spectral theorem there exists a unitary transformation to an orthonormal basis of eigenvectors of H. The previous parameters $\boldsymbol{a}$ may be expressed as linear combinations of rescaled parameters $\boldsymbol{z}$ in the eigenbasis. In this new basis, the neighborhood where the quadratic approximation is valid is now an N_p-sphere [13]:

$$\sum_{k=1}^{N_p} z_k^2 \leq T^2 \tag{4.6}$$

This defines a set of parameters $\{\boldsymbol{a}_i'\}$ lying on the boundary of the sphere. By an appropriate choice of T, these parameters represent a one-sigma error set of the model. Returning to the probability measure formalism, the expectation value of an observable $\mathcal{O}$ is just the observable evaluated with the minimized parameters:

$$\langle \mathcal{O}[d](\boldsymbol{x}) \rangle = \mathcal{O}[d(\boldsymbol{a}_0)](\boldsymbol{x}) \tag{4.7}$$

while the uncertainty associated with $\mathcal{O}$ is dependent on the one-sigma parameter set:

$$\sigma^2_{\mathcal{O}}(\boldsymbol{x}) = \sum_{i=1}^{N_p} \left\{\mathcal{O}[d(\boldsymbol{a}'_i)](\boldsymbol{x}) - \mathcal{O}[d(\boldsymbol{a}_0)](\boldsymbol{x})\right\}^2 \tag{4.8}$$

The magnitude of the eigenvalues of H describe how tightly constrained by the data an eigenparameter is. Large eigenvalues require only small changes in their respective parameters to vary $E[d]$ greatly, while small eigenvalues admit large changes in their parameters while having little effect on $E[d]$. Thus the Hessian method provide a direct analysis of the uncertainties of a distribution model. However it requires calculating second order partial derivatives of $E[d]$ with respect to parameters of a given model. For models with large numbers of parameters or complicated structure, performing Hessian analysis becomes difficult.

4.1.2 Monte Carlo Method

Numerically, the probability measure of the functional space may be determined by repeatedly sampling the space of data sets. A given data point has an associated uncertainty which describes the probability successive measurements return a value within some neighborhood of the data point. The measurements are assumed to sample a Gaussian distribution with matching uncertainty. The Monte Carlo method simulates successive measurements by randomly generating points based on their uncertainty. This creates an ensemble of measurements, similar to ensembles of systems in statistical physics.

The basis of Monte Carlo sampling begins with creating N_{rep} replicas of the data set $\{y_i\}$, which are denoted $\boldsymbol{y}^k = \{y_i^k\}$ for $k = 1, \ldots, N_{\text{rep}}$. Initially each replica is equally weighted, so for an arbitrary function g depending on the data set, its ensemble average over the replicas is defined as

$$\langle g \rangle_{\text{MC}} = \frac{1}{N_{\text{rep}}} \sum_{k=1}^{N_{\text{rep}}} g(\boldsymbol{y}^k) \tag{4.9}$$

The replicas are randomly generated in a Gaussian distribution such that the ensemble average of a data point recovers the data point itself, while the ensemble covariance recovers the experimental covariance matrix [5]:

$$\langle y_i \rangle_{\text{MC}} \approx y_i \qquad \langle y_i y_j \rangle_{\text{MC}} \approx y_i y_j + \text{cov}_{ij} \qquad \forall i, j = 1, \ldots, N_D \tag{4.10}$$

For every Monte Carlo replica k, define $E_k[d]$ to be the best-fit measurement which uses the kth replica as its experimental data set, and fit a distribution model d_k which minimizes E_k. This generates a Monte Carlo ensemble $\{d_k\}$ of distribution functions, which fully contains all of their associated uncertainty. For an arbitrary observable $\mathcal{O}$, the expectation value and standard deviation are given in terms of ensemble averages over $\{d_k\}$:

$$\begin{aligned} \mathcal{O}_0 = \langle \mathcal{O}[d] \rangle &= \langle \mathcal{O}[d_k] \rangle_{\text{MC}} \\ \sigma^2_{\mathcal{O}} &= \langle (\mathcal{O}_0 - \mathcal{O}[d_k])^2 \rangle_{\text{MC}} \end{aligned} \tag{4.11}$$

As shown in [14], the Monte Carlo method reproduces the Hessian method. By generating replicas in the parameter space (specifically, normally distributed in the eigenbasis of the parameter space), the one-sigma error set of a Hessian method may be approximated by a MC ensemble, and thus the method reproduced. The Monte Carlo method will be used throughout the rest of the paper.

4.2 Parameterization

Once an effective probability measure has been chosen for the distribution space, a model needs to be chosen. In general, models are chosen with the constraint that at $x = 1$, $d(1, Q^2) = 0$. In practice, the distribution is evaluated on a grid $\mathcal{G} = \{x_i\} \times \{Q_i\}$. First, an initial scale $Q_0^2 \in \{Q_i\}$ is chosen, and the distribution is evaluated on the grid line $\{d(x, Q_0^2) : x \in \{x_i\}\}$. Then the results are evolved to each energy scale Q_i via the DGLAP equations. Points not on the grid are interpolated using a suitable interpolation function, such as cubic splines (see section 2.1.2).

4.2.1 Analytic

Theoretical calculations predict that large and small x scale by a power law [8]. Thus the distribution may be given by the form [10]:

$$d_i(x, Q_0^2) = A_i x^{\alpha_i} (1 - x)^{\beta_i} \mathcal{F}(x, \boldsymbol{c}_i) \tag{4.12}$$

where A_i is a normalization parameter, α_i and β_i are power parameters, and $\boldsymbol{c}_i$ are parameters for the interpolation function $\mathcal{F}$ at an initial scale Q_0^2. The standard parameterization [10] of $\mathcal{F}$ is:

$$\mathcal{F}(x, \alpha_i, \beta_i, \delta_i, \gamma_i) = \frac{1 + \gamma_i (1 - x)^{\delta_i}}{B[2 + \alpha_i, \beta_i + 1] + \gamma_i B[2 + \alpha_i, \beta_i + \delta_i + 1]} \tag{4.13}$$

where $B[a, b]$ is the Euler beta function. By appropriate choice of parameters, the standard parameterization reduces to previous parameterizations used in earlier distribution function work [10] [39]. The standard parameterization assumes a functional form of the distribution function, which may produce artifacts in future predictions if the model substantially differs from the actual form.

Alternatively, complete orthogonal bases of the distribution function space may be used to approximate the interpolation function. Chebyshev polynomials have been used [14], but little work has been done with them due to computational complexity. In theory, other sets such as the Legendre polynomials may be used, but likely will have similar drawbacks. The benefit, however, is that there is no bias in approximating a functional form of the model.

4.2.2 Neural Network

In the early 90's, Cybenko and Hornik proved the universal approximation theorem [12], a corollary of which states that feed-forward neural networks are capable of approximat-

ing continuous functions between Euclidean spaces. Therefore, given appropriate weights and thresholds, a finite neural network is capable of acting as an unbiased interpolation function. Thus the problem of finding appropriate parameters for an interpolation function now reduces to training a neural network to reproduce a given function. Since f itself is a continuous function, it may be approximated directly by the neural network, and thus further avoid constraining by power law scaling.

For the purposes of this paper, a neural network may be considered to be just a mathematical function. Specifically, neural networks are multi-layered feed-forward perceptrons [5] with N_L layers and $\{n_i\}$ nodes per layer with $i = 1, \ldots, N_L$. An activation function $\boldsymbol{g}_i : \mathbb{R}^{n_{i-1}} \to \mathbb{R}^{n_i}$ is fixed for each layer $i > 1$. The network is dependent on a set of weights $\{\boldsymbol{w}_i\} \subset \mathbb{R}^{n_{i-1}}$ and thresholds $\{\boldsymbol{\theta}_i\} \subset \mathbb{R}^{n_i}$, again for layers $i > 1$. The weights and thresholds are the parameters of neural network which are varied to fit the desired distribution function. A neural network $\mathcal{N} : \mathbb{R}^{n_1} \to \mathbb{R}^{n_{N_L}}$ is thus defined recursively for an input $\boldsymbol{x} \in \mathbb{R}^{n_1}$:

$$\begin{aligned} \mathcal{N}(\boldsymbol{x}; \{\boldsymbol{w}_i\}, \{\boldsymbol{\theta}_i\}) &= \boldsymbol{l}_{N_L} \\ \boldsymbol{l}_i &= g_i(\boldsymbol{w}_i \cdot \boldsymbol{l}_{i-1} - \boldsymbol{\theta}_i) \\ \boldsymbol{l}_1 &= \boldsymbol{x} \end{aligned} \tag{4.14}$$

Fitting a neural network is done by minimizing a cost function, in this case the best-fit functional $E[d]$. Given an initial guess of weights and thresholds, the network is then trained by back-propagation algorithms which perform gradient descent of the cost function in the combined weight-threshold parameter space. There's an abundance of minimization algorithms, including stochastic gradient descent and ADAM [40], which are beyond the scope of this paper, but mainly differ by speed and efficiency of reaching a minimum. It shall be assumed that all reach an appropriate estimate of the global minimum.

To ensure that $f(x, Q_0^2) = 0$ as $x \to 1$, the evaluation of the network at $x = 1$ is subtracted from all x values:

$$\mathcal{N}'(x, Q_0^2; \{\boldsymbol{w}_i\}, \{\boldsymbol{\theta}_i\}) = \mathcal{N}(x, Q_0^2; \{\boldsymbol{w}_i\}, \{\boldsymbol{\theta}_i\}) - \mathcal{N}(1, Q_0^2; \{\boldsymbol{w}_i\}, \{\boldsymbol{\theta}_i\}) \tag{4.15}$$

The application of neural networks to fit distribution functions is currently being applied by the NNPDF Collaboration to create PDF sets [11]. The Collaboration's efforts have been successful in reproducing various high-energy phenomena at the LHC [11]. Similar work has been done for unpolarized fragmentation functions of the pion by the MAP Collaboration. Their `MAPFF` set will be used for reweighting later in this paper [5].

4.3 Reweighting

Distribution functions are reported as a set of Monte Carlo replicas of points fitted against sets of experimental data. However, experiments regularly produce new data sets. The most direct way of fitting this new data is to refit the distribution entirely with both the old and new data. For complicated fitting procedures, refitting is laborious and is rather

done when techniques improve or large bulk data becomes available. Some probability analysis allows refitting to be avoided entirely.

Consider a distribution function $f(x, Q^2)$, which has been fitted against a previous data set A (the notation of PDFs is used to avoid confusion with the differential d). The distribution has an associated probability $\mathcal{P}(f|A)\,\mathrm{D}f$, which is represented as an ensemble of Monte Carlo replicas $\{f_k\}$ for $k = 1, \ldots, N_{\text{rep}}$. The probability is used to compute observables $\mathcal{O}[f]$ via the ensemble average (4.9):

$$\langle \mathcal{O}[f] \rangle_A = \int \mathcal{O}[f] \mathcal{P}(f|A)\,\mathrm{D}f = \frac{1}{N_{\text{rep}}} \sum_{k=1}^{N_{\text{rep}}} \mathcal{O}[f_k] \tag{4.16}$$

where $\langle \cdot \rangle_A$ denotes an average over distributions fitted against A. Suppose a new data set B is to be included with A giving the total data set AB. Computing observables with the combined data set amounts to determining $\mathcal{P}(f|AB)$. Such a process is called reweighting, since the end result appears as weights of A's ensemble average. The following proof the reweighting process is adapted from [15].

Let B be a data set consisting of n points, or equivalently as a vector $B \in \mathbb{R}^n$. Probability distributions P may be defined for probability measures $\mathcal{P}$ on a local volume of their respective spaces, i.e.

$$\begin{aligned} P(f|A) &= \mathcal{P}(f|A)\,\mathrm{D}f \\ P(B|fA) &= \mathcal{P}(B|fA)\,\mathrm{d^n}B \end{aligned}$$

Applying Bayes' theorem to probability distributions combined with the above allows the theorem to be applied to probability measures:

$$\begin{aligned} P(fB|A) =& P(f|AB)\,P(B|A) = P(B|fA)\,P(f|A) \\ \implies & \mathcal{P}(f|AB)\,\mathrm{D}f\,\mathcal{P}(B|A)\,\mathrm{d^n}B = \mathcal{P}(B|fA)\,\mathrm{d^n}B\,\mathcal{P}(f|A)\,\mathrm{D}f \end{aligned}$$

To calculated the desired quantity $\mathcal{P}(f|AB)\,\mathrm{D}f$, the quantities $P(B|A)$ and $P(B|fA)$ must be well defined in their local neighborhoods to take the following ratio:

$$\mathcal{P}(f|AB)\,\mathrm{D}f = \frac{\mathcal{P}(B|fA)\,\mathrm{d^n}B}{\mathcal{P}(B|A)\,\mathrm{d^n}B} \mathcal{P}(f|A)\,\mathrm{D}f \tag{4.17}$$

However, taking the limit as the volume $\mathrm{d^n}B$ goes to zero is ambiguous. To treat the neighborhoods consistently, they will be converted to n-spheres where the volume vanishes only as the radial coordinate vanishes, thus making the zero-volume limit explicit. Note in the above that there is no dependence of the denominator $P(B|A)$ on f. Requiring that the probability measure $\mathcal{P}(f|AB)$ is normalized fixes the denominator:

$$P(B|A) = \int P(B|fA)\,\mathcal{P}(f|A)\,\mathrm{D}f \tag{4.18}$$

This is precisely an ensemble average over A, as defined in (4.16):

$$P(B|A) = \langle P(B|fA) \rangle_A = \frac{1}{N_{\text{rep}}} \sum_{k=1}^{N_{\text{rep}}} P(B|f_k A) \tag{4.19}$$

Or in terms of the probability measure:

$$\mathcal{P}(B|A)\,\mathrm{d^n}B = \frac{1}{N_{\text{rep}}}\sum_{k=1}^{N_{\text{rep}}}\mathcal{P}(B|f_kA)\,\mathrm{d^n}B \tag{4.20}$$

The measure $\mathcal{P}(B|f_kA)\,\mathrm{d^n}B$ is the probability that a given f_k lies in a $\mathrm{d^n}B$ neighborhood centered about B. Assuming that all uncertainties are Gaussian, the density is given by the multivariate Gaussian distribution

$$\mathcal{P}(B|fA) = \frac{e^{-\frac{1}{2}\chi^2[f;B]}}{\sqrt{(2\pi)^n\det(\mathrm{cov}_B)}} \tag{4.21}$$

where $\chi^2[f;B]$ is the chi-squared best fit (4.2) with respect to B's data, while cov_B is B's covariance matrix. To convert this to n-spherical coordinates, perform a change of basis to B' which maps cov_B to the identity matrix (this is always possible by the spectral theorem, since a covariance matrix is real, symmetric, and has positive eigenvalues), and translate to be centered on f. The relevant changes are

$$\mathrm{d^n}B = \sqrt{\det(\mathrm{cov}_B)}\,\mathrm{d^n}B' \qquad \text{and} \qquad \chi^2[f;B] = \sum_{i=1}^{n}(B')_i^2 \tag{4.22}$$

Notice that χ takes the form of a radial component with respect to a rectangular space B', and thus converting to spherical coordinates in n-dimensions:

$$\mathrm{d^n}B' = \frac{2\pi^{n/2}}{\Gamma(n/2)}\chi^{n-1}\,\mathrm{d}\chi\,\mathrm{d^{n-1}}\Omega \tag{4.23}$$

Where $\mathrm{d^{n-1}}\Omega$ is the standard solid angle measure. Thus the multivariate Gaussian distribution becomes

$$\mathcal{P}(B|fA)\,\mathrm{d^n}B = \frac{2^{1-n/2}\chi^{n-1}[f;B]}{\Gamma(n/2)}e^{-\frac{1}{2}\chi^2[f;B]}\,\mathrm{d}\chi\,\mathrm{d^{n-1}}\Omega \tag{4.24}$$

This expresses the numerator of (4.17) in terms of spherical χ coordinates. In turn, (4.24) defines the probability of measuring $\chi[f;B]$ given a function f fitted against A:

$$\mathcal{P}(\chi|fA)\,\mathrm{d}\chi = \frac{2^{1-n/2}\chi^{n-1}}{\Gamma(n/2)}e^{-\frac{1}{2}\chi^2}\,\mathrm{d}\chi \tag{4.25}$$

As with $P(B|fA)$, the χ probability is an ensemble average over A:

$$\begin{aligned}\mathcal{P}(\chi|A) &= \int\mathcal{P}(\chi|fA)\,\mathcal{P}(f|A)\,\mathrm{D}f = \frac{1}{N_{\text{rep}}}\sum_{k=1}^{N_{\text{rep}}}\mathcal{P}(\chi_k|f_kA)\\ &= \frac{2^{1-n/2}}{\Gamma(n/2)N_{\text{rep}}}\sum_{k=1}^{N_{\text{rep}}}\chi_k^{n-1}e^{-\frac{1}{2}\chi_k^2}\end{aligned} \tag{4.26}$$

Combining equations (4.20) and (4.26) expresses the denominator of (4.17) in terms of χ as well:

$$\mathcal{P}(B|A)\,\mathrm{d^n}B = \langle\mathcal{P}(B|fA)\rangle_A\,\mathrm{d^n}B = \langle\mathcal{P}(\chi|fA)\rangle_A\,\mathrm{d}\chi\,\mathrm{d^{n-1}}\Omega = \mathcal{P}(\chi|A)\,\mathrm{d}\chi\,\mathrm{d^{n-1}}\Omega \tag{4.27}$$

Returning to equation (4.17) with the new χ measures:

$$\begin{aligned}\mathcal{P}(f|AB)\,\mathrm{D}f &= \frac{\mathcal{P}(\chi|fA)\,\mathrm{d}\chi\,\mathrm{d}^{\mathrm{n}-1}\Omega}{\mathcal{P}(\chi|A)\,\mathrm{d}\chi\,\mathrm{d}^{\mathrm{n}-1}\Omega}\mathcal{P}(f|A)\,\mathrm{D}f \\ &= \frac{\mathcal{P}(\chi|fA)}{\mathcal{P}(\chi|A)}\mathcal{P}(f|A)\,\mathrm{D}f\end{aligned} \tag{4.28}$$

Since $\mathrm{d}^{\mathrm{n}-1}\Omega$ is a solid angle, it is non-vanishing in the zero-volume limit, and thus directly cancels. Since $\mathrm{d}\chi$ is one-dimensional, it approaches zero in the zero-volume limit, but does so in the same neighborhood for both the numerator and denominator. The zero-volume limit is therefore well-defined, and $\mathrm{d}\chi$ may be canceled as well. Now (4.28) may be substituted into $\langle\mathcal{O}[f]\rangle_{AB}$ as defined by (4.16):

$$\begin{aligned}\langle\mathcal{O}[f]\rangle_{AB} &= \int \mathcal{O}[f]\mathcal{P}(f|AB)\,\mathrm{D}f \\ &= \int \mathcal{O}[f]\frac{\mathcal{P}(\chi|fA)}{\mathcal{P}(\chi|A)}\mathcal{P}(f|A)\,\mathrm{D}f \\ &= \left\langle \mathcal{O}[f_k]\frac{\mathcal{P}(\chi|f_kA)}{\mathcal{P}(\chi|A)}\right\rangle_A \\ &= \frac{1}{N_{\mathrm{rep}}}\sum_{k=1}^{N_{\mathrm{rep}}}\frac{\mathcal{P}(\chi|f_kA)}{\mathcal{P}(\chi|A)}\mathcal{O}[f_k]\end{aligned} \tag{4.29}$$

The final expression takes the form a weighted average, where the weights $\{w_k\}$ are

$$w_k = \frac{\mathcal{P}(\chi|f_kA)}{\mathcal{P}(\chi|A)} \qquad \text{with} \qquad \sum_{k=1}^{N_{\mathrm{rep}}} w_k = N_{\mathrm{rep}} \tag{4.30}$$

Notice that by application of Bayes theorem again, the weights may be written

$$w_k = \frac{\mathcal{P}(f_k|\chi A)}{\mathcal{P}(f_k|A)} = N_{\mathrm{rep}}\mathcal{P}(f_k|\chi A) \tag{4.31}$$

which promotes the interpretation that the reweighting process weights each replica by the probability it is generated by the new combined dataset. Weights may be directly calculated from (4.25) and (4.26):

$$w_k = \frac{(\chi_k^2)^{(n-1)/2}e^{-\frac{1}{2}\chi_k^2}}{\frac{1}{N_{\mathrm{rep}}}\sum_{i=1}^{N_{\mathrm{rep}}}(\chi_i^2)^{(n-1)/2}e^{-\frac{1}{2}\chi_i^2}} \tag{4.32}$$

As (4.32) shows, reweightings are entirely determined by the $\chi_k^2 = \chi^2[f_k;B]$ for each replica f_k. Because χ^2 takes such a central role in reweighting, specifically from multivariate Gaussian distributions in (4.21), it's naturally chosen to be the best-fit function.

4.3.1 Reweighting Multiple Experiments

Consider reweighting a distribution function ensemble $\{f_k\}$ by two independent experiments B_1 and B_2 simultaneously. Because the experiments are independent, the covari-

ance matrix factorizes into the covariance matrices of the individual experiments:

$$\text{cov} = \left(\begin{array}{c|c} \text{cov}_1 & 0 \\ \hline 0 & \text{cov}_2 \end{array}\right) \qquad \text{and} \qquad \text{cov}^{-1} = \left(\begin{array}{c|c} \text{cov}_1^{-1} & 0 \\ \hline 0 & \text{cov}_2^{-1} \end{array}\right) \tag{4.33}$$

The χ^2 similarly factorizes:

$$\begin{aligned} \chi^2 &= \left(\begin{array}{c|c} \Delta \boldsymbol{y}_1^T & \Delta \boldsymbol{y}_2^T \end{array} \right) \left(\begin{array}{c|c} \text{cov}_1^{-1} & 0 \\ \hline 0 & \text{cov}_2^{-1} \end{array}\right) \left(\begin{array}{c} \Delta \boldsymbol{y}_1 \\ \hline \Delta \boldsymbol{y}_2 \end{array}\right) \\ &= \Delta \boldsymbol{y}_1^T \text{cov}_1^{-1} \Delta \boldsymbol{y}_1 + \Delta \boldsymbol{y}_2^T \text{cov}_2^{-1} \Delta \boldsymbol{y}_2 \\ &= \chi_1^2 + \chi_2^2 \end{aligned} \tag{4.34}$$

Directly substituting into (4.32) gives [15]:

$$w_k \propto (\chi_{1;k}^2 + \chi_{2;k}^2)^{(n_1+n_2-1)/2} e^{-\frac{1}{2}(\chi_{1;k}^2+\chi_{2;k}^2)} \tag{4.35}$$

again normalized such that the sum is N_{rep}. This process may be repeated for a finite number of data sets $\{B_i\}$ for $i = 1, \ldots, N$:

$$\begin{aligned} w_k &\propto (\chi_k^2)^{(n-1)/2} e^{-\frac{1}{2}\chi_k^2} \\ \chi_k^2 &= \sum_{i=1}^{N} \chi_{i;k}^2 \\ n &= \sum_{i=1}^{N} n_i \end{aligned} \tag{4.36}$$

Now instead, consider reweighting an already reweighted ensemble $\{f_k' = w_{1;k} f_k\}$ by a second data set B_2. The final weights should match (4.35) and may be directly calculated by

$$w_k' = \frac{w_{1+2;k}}{w_{1;k}} = \frac{(\chi_{1;k}^2 \chi_{2;k}^2)^{(n_1+n_2-1)/2}}{(\chi_{1;k}^2)^{(n-1)/2}} e^{-\frac{1}{2}\chi_{2;k}^2} \tag{4.37}$$

The weights are not given by $\{w_{2;k}\}$ on the original $\{f_k\}$ since the probability $\mathcal{P}(f|\chi_1\chi_2)$ does not factorize: $\chi_{2;k}^2$ is dependent on $f_k' = w_{1;k} f_k$ and is thus dependent on the initial reweighting $w_{1;k}$. For independent experiments the order of reweighting is irrelevant since the results of either order is guaranteed to match (4.35). However, this involves determining $\chi_{1;k}^2$ for the previous experiment, which might not be readily available.

4.4 Unweighting

Reporting a reweighted distribution ensemble is unfavorable as it's inefficient for large ensembles, requires adapting existing code bases, and makes successive reweighting difficult. Reweighted ensembles may be converted to an equally weighted ensemble via unweighting [41]. A description of the unweighting method is given here; for an alternative derivation see [41].

To unweight a reweighted ensemble $\{f_k; w_k\}$ for $k = 1, \ldots, N_{\text{rep}}$, fix a new ensemble size N'_{rep}. For each weight, define the probability cummulant P_k as the normalized sum of all weights before k:

$$P_k = \sum_{j=1}^{N_{\text{rep}}} \frac{w_j}{N_{\text{rep}}}, \qquad P_0 = 0 \tag{4.38}$$

The cummulant P_0 is defined to be zero for convenience. This defines a partitioning of the unit interval into sub-intervals $\{[P_{k-1}, P_k] : k = 1, \ldots, N_{\text{rep}}\}$ where kth sub-interval has a size of w_k/N_{rep}. A similar partitioning may be performed on the new ensemble. Since each replica in this ensemble has an equal weight of one, the partitioning of the unit interval is just $\{[r_{j-1}, r_j] : j = 1, \ldots, N'_{\text{rep}}]\}$ for $r_j = j/N'_{\text{rep}}$. For each reweighted replica, define its unweighting as:

$$w'_k = \sum_{j=1}^{N'_{\text{rep}}} \theta\left(r_j - P_{k-1}\right) \theta\left(P_k - r_j\right) \tag{4.39}$$

and $\theta(x)$ is the Heaviside theta function. Note that w'_k is a non-negative integer. Equation (4.39) is interpreted as the number of copies of the reweighted replica f_k to make in the unweighted ensemble to reproduce the weighting w_k. Each $1/N'_{\text{rep}}$ interval corresponds to a single replica f'_j in the unweighted ensemble. In equation (4.39), a replica f'_j is a copy of f_k only when the right-hand side of f'_j's interval lies in f_k's interval. That is, when $P_{k-1} < j/N'_{\text{rep}} \leq P_k$, which is precisely the summand term. Thus w'_k indeed counts the number of copies of f_k to be made. This ensures that the w'_k satisfy the correct normalization of a N'_{rep} ensemble:

$$\sum_{k=1}^{N_{\text{rep}}} w'_k = N'_{\text{rep}} \tag{4.40}$$

Unweighting preserves the probability density for large enough N'_{rep}, since

$$p'_k = \frac{w'_k}{N'_{\text{rep}}} \to \frac{w_k}{N_{\text{rep}}} = p_k \tag{4.41}$$

in the limit of $N'_{\text{rep}} \to \infty$. This happens because smaller unweighted intervals allow for more accurate sampling of the original partitioning, and approach perfect sampling as the interval size vanishes.

4.4.1 Replica Ensemble Information Measurement

Reweighting an ensemble changes the importance of each replica in calculating observables. Replicas with relatively small weights will give negligible contributions to the average, and thus may be removed from a replica set. The information loss of reweighting may be quantified via information entropy. Given a distribution function ensemble $\{f_k\}$ with a probability density $\mathcal{P}(f_k)$, the information a replica f_k carries may be defined by the Shannon information:

$$I(\mathcal{P}(f_k)) = \ln\left(\frac{1}{\mathcal{P}(f_k)}\right) = -\ln \mathcal{P}(f_k) \tag{4.42}$$

To characterize the information distribution in a set, the Shannon entropy of $\{f_k\}$ is given by the ensemble average of the information:

$$H(\{f_k\}) = \langle I(\mathcal{P}(f))\rangle_{\{f_k\}} = -\sum_{k=1}^{N_{\text{rep}}} \mathcal{P}(f_k) \ln \mathcal{P}(f_k) \tag{4.43}$$

For an equally weighted ensemble, such as an initially fitted distribution ensemble, the probability of a replica f_k is just $\mathcal{P}(f_k) = 1/N_{\text{rep}}$, so the entropy is minimized:

$$H = -\sum_{k=1}^{N_{\text{rep}}} \left(\frac{1}{N_{\text{rep}}}\right) \ln \left(\frac{1}{N_{\text{rep}}}\right) = \ln N_{\text{rep}} \tag{4.44}$$

For a reweighted ensemble, the probability is given by the weights in equation (4.31), and has an entropy of

$$H = -\sum_{k=1}^{N_{\text{rep}}} \left(\frac{w_k}{N_{\text{rep}}}\right) \ln \left(\frac{w_k}{N_{\text{rep}}}\right) = \frac{1}{N_{\text{rep}}} \sum_{k=1}^{N_{\text{rep}}} w_k \ln(N_{\text{rep}}/w_k) \tag{4.45}$$

Matching the entropy of a reweighted ensemble with N_{rep} replicas (4.45) to an ensemble with N_{eff} equally weighted replicas (4.44) gives the effective number of replicas:

$$N_{\text{eff}} = \exp\left(\frac{1}{N_{\text{rep}}} \sum_{k=1}^{N_{\text{rep}}} w_k \ln(N_{\text{rep}}/w_k)\right) \tag{4.46}$$

Thus a reweighted ensemble carries the same effective information as a N_{eff} fitted ensemble. When unweighting a reweighted ensemble, it is therefore appropriate to use (4.46) as the unweighted ensemble size, since the addition of more replicas will not carry further information, while the removal of any replica will lose information.

Chapter 5

Results and Discussion

This chapter describes the results of numerically calculating differential cross sections of unpolarized hadron-hadron scattering and the consequent reweighting of the summed pion fragmentation functions. All theoretical data was calculated via unpolarized hadron-hadron code to $O(\alpha_S^3)$ with the code of [42] [43]. The `MAPFF1.0NLOPIsum` FF set and `NNPDF3.1_nlo_as_0118` PDF set were used for all calculations, and $\mu_{\text{ren}} = \mu_{\text{fact}} = \mu_{\text{frag}}$ was chosen for all processes. Plots showing theoretical observables (including plots of FF sets) display a line representing the central value calculated by the ensemble average (4.9) centered in a band whose width is the standard deviation given by (4.11). Plots of data points display the uncertainty as the uncorrelated error summed in quadrature, but do not show the normalization error. All FFs are plotted at $\mu = 5$ GeV unless otherwise stated.

Reweighting, as outlined in section 4.3, was performed on the total experimental data sets. Each data set was filtered against a momentum cutoff of $p_T^{\min} = 5$ GeV. An ensemble of observables was calculated for each experiment by sampling the FF replicas. For every FF replica $D_k^{\pi^0}$, the $\chi^2_{i;k}$ value for the ith experiment was calculated and summed over all experiments. The reweight was computed and normalized via equation (4.36). This defines the reweighted FF set described in this chapter.

5.1 Data-Theory Comparison

Figures 5.1, 5.2, and 5.3 show the data-theory comparison of the filtered data sets to the original and reweighted FFs. With respect to the observables, the reweighting does not significantly alter the central value. The only appreciable change is a tightening of variance in the ALICE data sets in the high-p_T range (> 20 GeV). Table 5.1 shows the effect of reweighting on the χ^2 fits of individual experiments, as well as the total combined fit. While reweighting is not guaranteed to maintain or lower the χ^2 of individual fits, it does ensure the total χ^2 is minimized. Note that sets which have an increased χ^2 in total contain 26 points, which is $\sim 1/4$ of the total 101 points used in reweighting. Due to the small size of the increased χ^2 sets, whether reweighting is an actual decrease in accuracy on these sets is not conclusive.

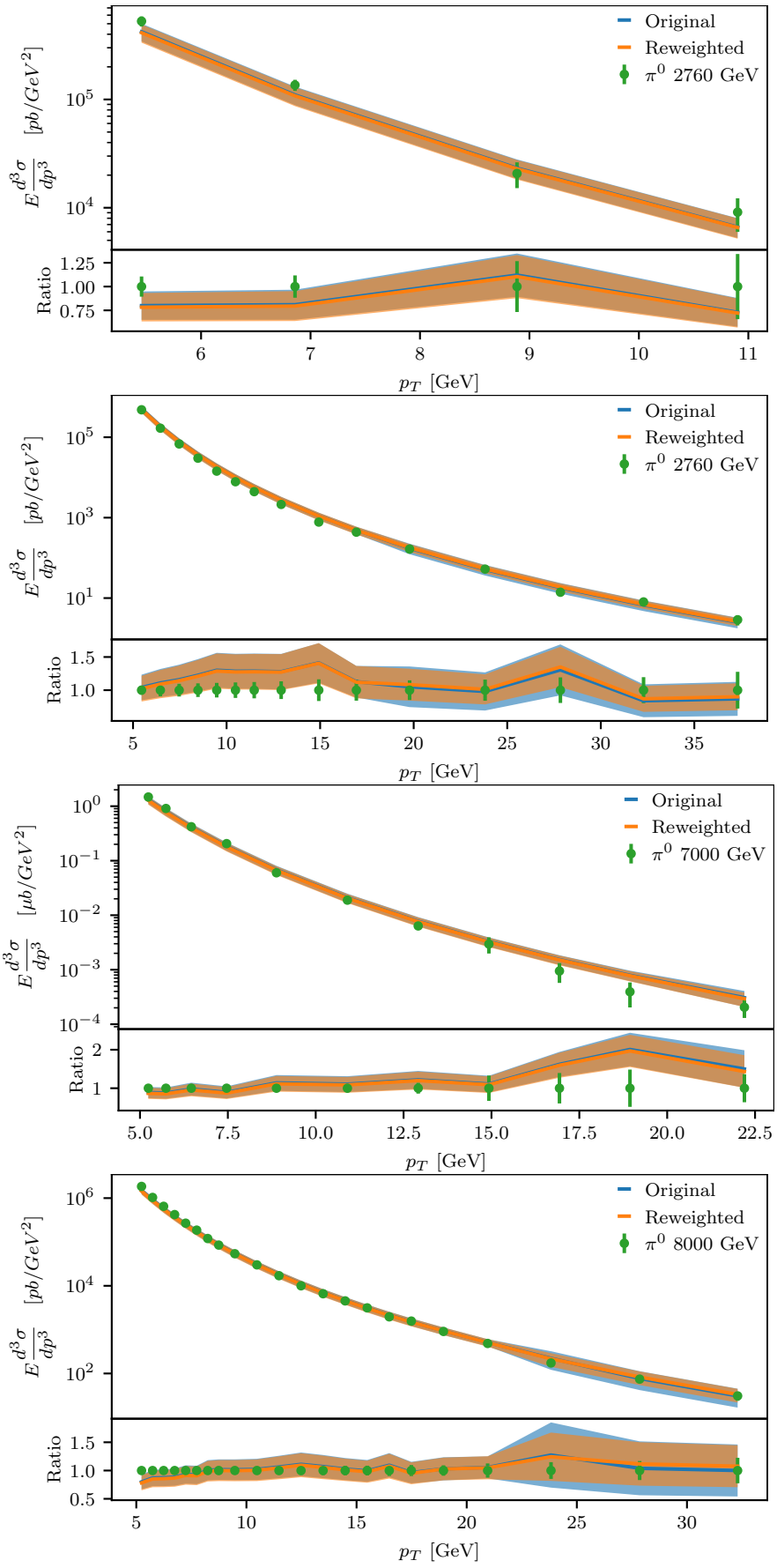

Figure 5.1: Comparison of original and reweighted FF sets on ALICE data.

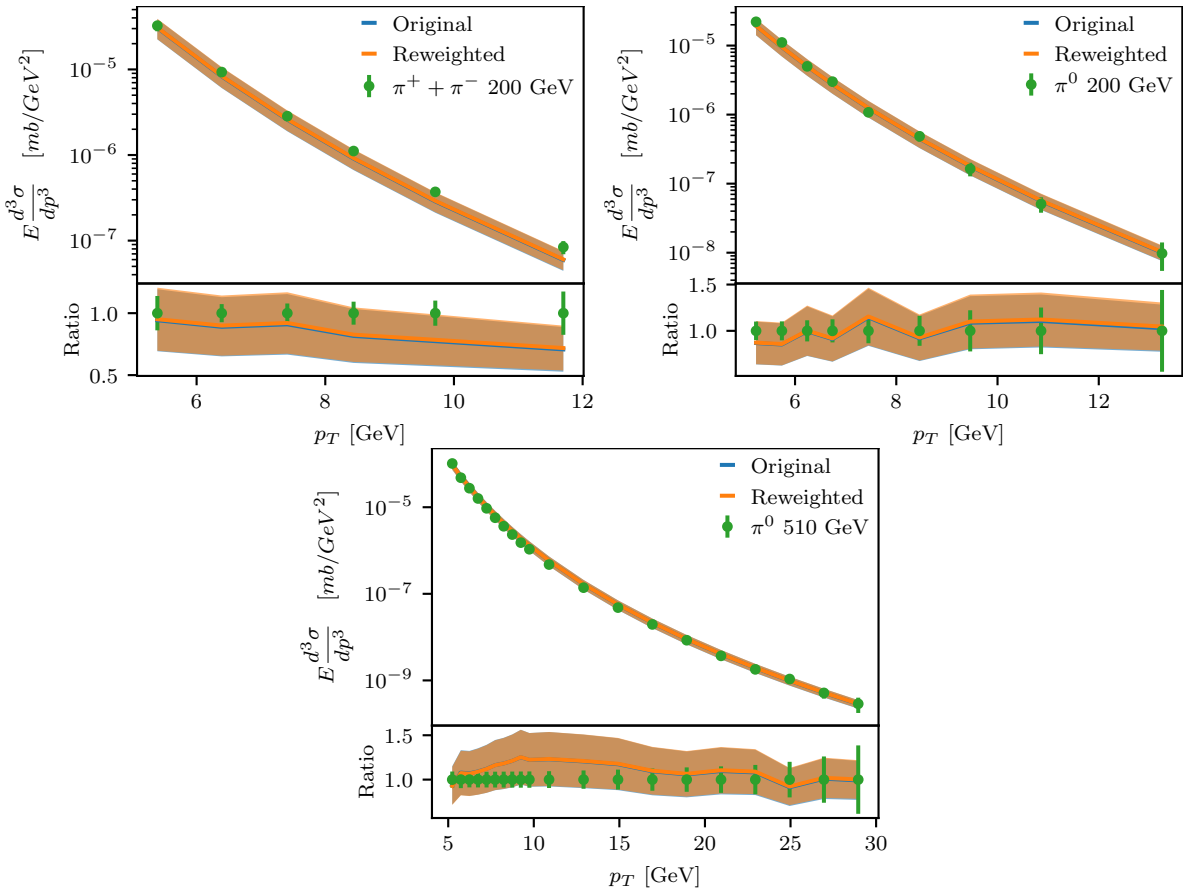

Figure 5.2: Comparison of original and reweighted FF sets on PHENIX data.

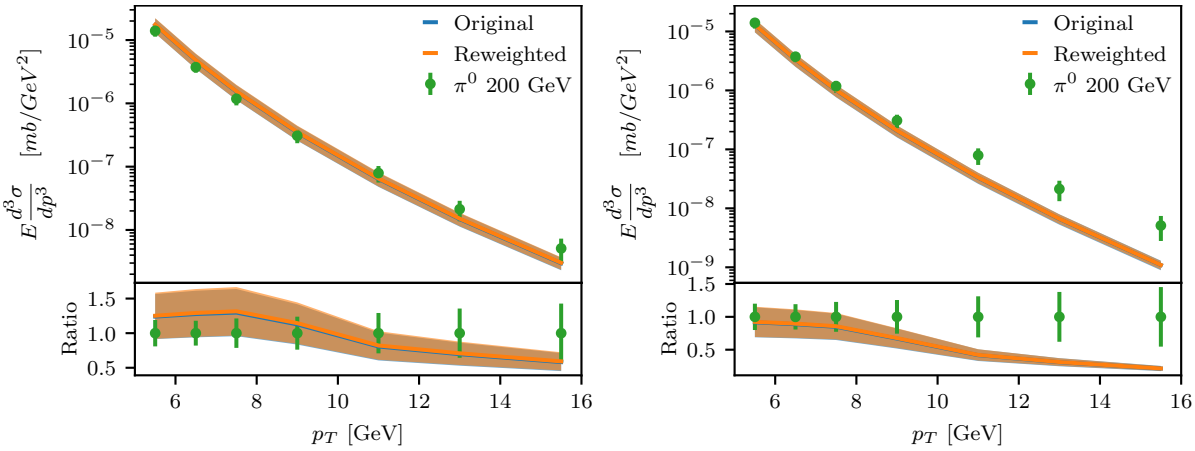

Figure 5.3: Comparison of original and reweighted FF sets on STAR data.

Experiment	Original χ^2/point	Reweighted χ^2/point	N_{dat} filtered
ALICE 2760 π^0 [24]	1.45	2.00	4
ALICE 2760 π^0 [25]	1.53	0.97	15
ALICE 7000 π^0 [28]	2.42	1.87	11
ALICE 8000 π^0 [30]	1.52	1.36	22
ALICE Total	1.71	1.49	52
PHENIX 200 $\pi^+ + \pi^-$ [32]	0.75	1.29	6
PHENIX 200 π^0 [33]	0.56	0.67	9
PHENIX 510 π^0 [34]	0.75	0.48	20
PHENIX Total	0.70	0.67	35
STAR 200 π^0 [36]	0.98	0.64	7
STAR 200 π^0 [37]	1.38	1.68	7
STAR Total	1.18	1.16	14
Total	**1.29**	**1.12**	**101**

Table 5.1: χ^2 per point for each experiment for both the original and reweighted sets. The cumulative χ^2 per point for each set of experiments is also listed.

The total χ^2 per point is 1.12, which indicates the reweighted FF fits the data well. The ALICE experiments contain the most fitted data points, and, while acceptable, they also had the worst overall fit at 1.49. The total lower energy results $\sqrt{s} < 1$ TeV of PHENIX and STAR have a χ^2 per point of only 0.81, indicating the fit is indeed poorer for energies above the 1 TeV scale. The previous fitting of `MAPFF` was performed on SIA and SIDIS data on a scale of 10 to 100 GeV, while the range of new *pp*-collision data begins at 200 GeV. The successful predictions of `MAPFF` verifies the accuracy of the time-like DGLAP equations in describing the energy scale dependence of fragmentation functions. Since gluon partonic processes dominate the kinematic region below 20 GeV at high energies, the lack of change in this region indicates previous fits accurately described the gluon FFs. Rather, the high-p_T constraints correspond to constraints of the quark FF.

5.2 Reweighted Fragmentation Functions

The original FF set and reweighted FF set are plotted against each other in figure 5.4. Because only the summed pion distribution was reweighted, quarks and antiquarks are equivalent under charge conjugation and isospin symmetry. Thus only the real quarks and gluons are shown.

Quark fragmentation functions are highly constrained in the high-z region ($z \gtrsim 0.2$). Reweighting is unable to alter FFs in this region, so any dependence upon high-z behavior of quark FFs in *pp*-collisions that differs from SIA or SIDIS will appear as a discrepancy, even in the limit of large replica ensembles. Because of how well the fit describes the data,

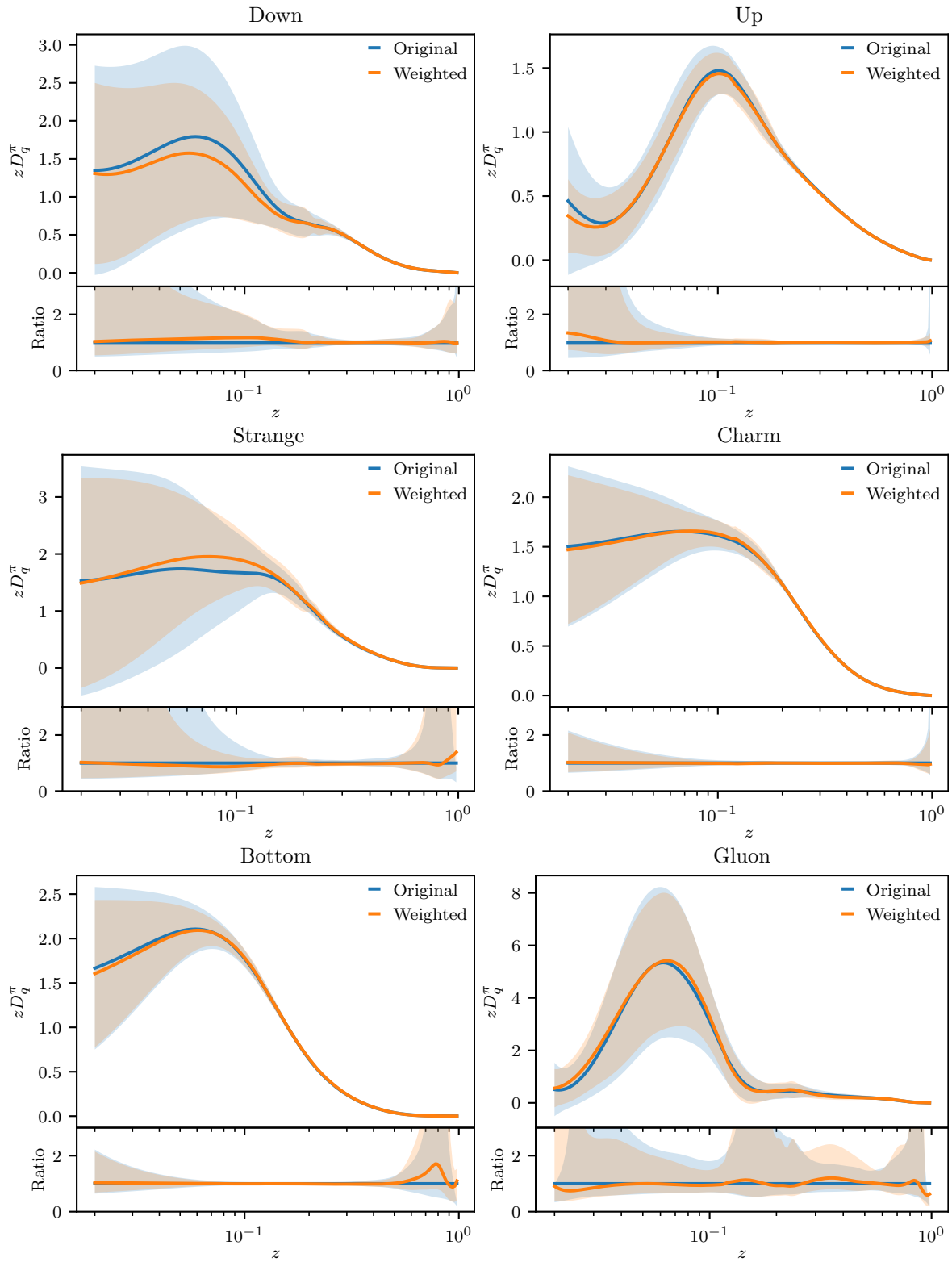

Figure 5.4: Comparison of original and reweighted FF sets.

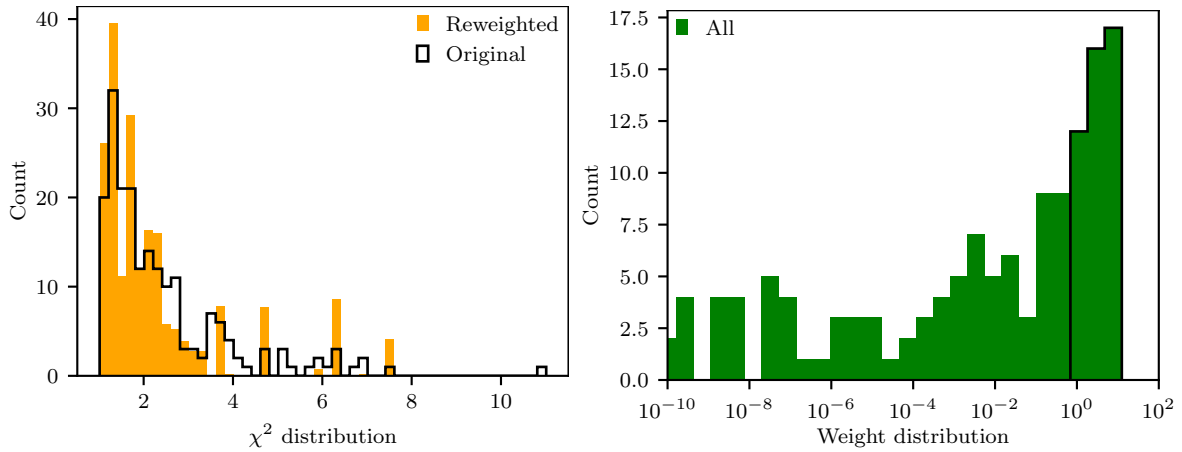

Figure 5.5: Left: Distribution of χ^2 values for each ensemble replica. Right: Histogram of reweighting weights. The outlined box indicates the weights corresponding to the highest contributing replicas.

the pp-collisions are unlikely to introduce such changes. Furthermore, the improvement of the χ^2 by reweighting indicates below the high-z region has a significant effect on fitting pp-collision data.

The sea quarks c, b exhibit small change, and only in the variance near low-z values. Although a sea quark, the s distribution shows a positive adjustment in the $0.05 \lesssim z \lesssim 0.2$ range. Most notably, the valence d quark shows the greatest change after reweighting. The central value differs for $z \leq 0.2$, and there's any accompanying constraint on the variance. While also a valence quark, u only shows a deviation in the low-z region.

The gluon fragmentation function is of special interest in pp-collisions due to its large relative contribution to partonic cross sections. Unlike the quark FFs, the gluon does not share the same constraints in the high-z region. Any high-z dependence of pp-collisions is therefore entirely contained in the gluon FF in this region. Reweighting had the largest effect on the gluon FF at high-z, while below this region it showed the same general constraining as the quark FFs. The lack of change in the gluon FF indicates that the previous SIA and SIDIS data were able to accurately describe the gluon's behavior. Surprisingly, reweighting has shown that the pp-collisions introduce constraints mainly into the quarks FFs.

5.3 Reweighting and Unweighting

The reweighting process may be characterized by the weight distribution shown in figure 5.5. As expected, the weight distribution roughly mirrors the χ^2_k distribution of replica fits in the original data-theory comparison. After reweighting, the χ^2 distribution contracts towards $\chi^2 = 1$, indicating the reweighted set improves the fit on a per replica basis. A measure of consistency $\mathcal{P}(\alpha)$ is tested in appendix B. This test confirms that the experimental uncertainties are well estimated by the reweighting.

Both the original and reweighted distributions are highly compatible. Overall there's only

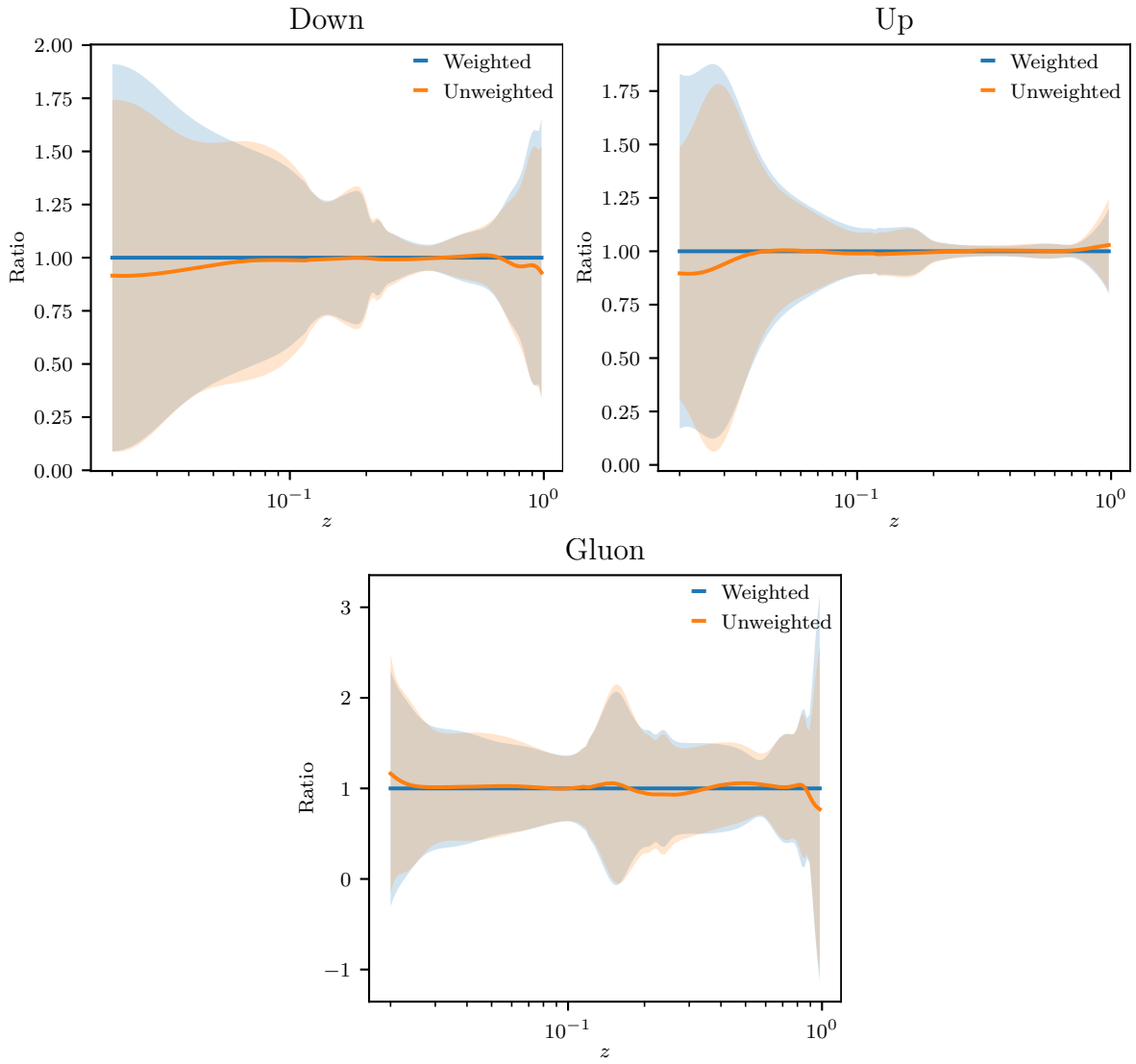

Figure 5.6: Sample ratios of unweighted and weighted FFs.

small deviations which are entirely contained in the variance of the original FF set. A more quantitative analysis may be performed through calculating the distances between the central values and the variances, which is done in appendix A. Comparing distances confirms that the FF sets are generated by the same underlying physical distribution.

To prepare the reweighted FF set for reporting, it was unweighted according to the unweight algorithm. The number of effective replicas as determined by [41] was $N_{\text{eff}} = 39$, which is shown in figure 5.5 as the outlined weights. These weights take values in the range $[0.6, 15]$. Because the probability distribution is expected to be preserved by unweighting, there should be little variation between the reweighted and unweighted distributions. Figure 5.6 shows the ratio of reweighted and unweighted FF sets. The changes in the unweighted central values is well contained in the variance of the FF sets, and the variances are roughly equivalent. While the unweighted ensemble contains 39 replicas, only 33 unique replicas of the original ensemble were preserved. The remaining 6 unweighted replicas are copies of the 33 unique ones in order to maintain the reweighting

probability distribution. As with the comparison of original and reweighted sets, the distances between the reweighted and unweighted sets is computed in appendix A, which provides the same conclusion.

The best test of reweighting is to perform an independent fit of the data. In the absence of a refit, reweighting acts as indicator of when a new fit would be productive. As shown previously, the *pp*-collision data is well reproduced by the previous SIA and SIDIS fit, outside the high energy and high-p_T region. When substantial data becomes available in such a region, it would then be advantageous to do a new fit to independently verify the results of reweighting.

5.4 Impact of Theoretical Choices

The choice of PDF set greatly contributes to *pp*-collision cross sections since they are sampled twice. Since `MAPFF` is a neural-network based fit, `NNPDF` is a natural choice of PDF that also provides a high-quality fit. The PDF variance was propagated through to the observables by randomly choosing a PDF replica for every FF replica. Because of time constraints, the effects of PDF variance were not quantified. These calculations consist of recalculating the observables and reweighting using only the central PDF value, as well as repeating the process for other PDF sets. This technique was applied to `MAPFF` [5] and `NNFF` [4] fits, which both showed little dependence upon the PDF variance. Because PDF sets are fitted against a larger independent data set (i.e. fully exclusive data), they are better determined than FF sets, so it's expected that PDF variance will have little effect on FF variance.

The FF fit is dependent upon the choice of cutoff $p_T^{\min}$ as it directly determines the number of data points used in the fit. While increasing cutoff improves the agreement, less data points are included in the reweighting, and thus provide a weaker constraint on the fragmentation functions. This dependence is shown in figure 5.7, where the χ^2/point measures the data-theory agreement while the total N_{dat} shows the decreasing number of data points available to constrain the FF. All choices of cutoff for 5 GeV $\leq p_T^{\min} \leq$ 10 GeV showed improvement of the total χ^2, and all reweighted χ^2/point were in an acceptable range of $\leq$ 1.2. The minimum cutoff possible is 2 GeV, which was the cutoff of the previous `MAPFF1.0NLOPI` fit, but for cutoffs smaller than 5 GEV the predictions began showing significant disagreement with data. This is in part due to numerical instabilities in the code for calculating hadron-hadron cross sections in the $<$ 5 GeV range. A cutoff of 5 GeV was chosen for the fits used in this paper since it encapsulates the largest kinematic range while maintaining a decent fit.

The influence of different experimental sets may be shown by performing separate reweightings on individual experimental subsets. Figure 5.8 shows the ratio of reweighting by LHC (ALICE) and non-LHC (PHENIX, STAR) versus the total reweighted FF set. The LHC data shows the strongest constraint on the variance of the FF, but the central values show significant disagreement with the non-LHC data. This effect may be due to the greater abundance and improved uncertainty bounds of the LHC data.

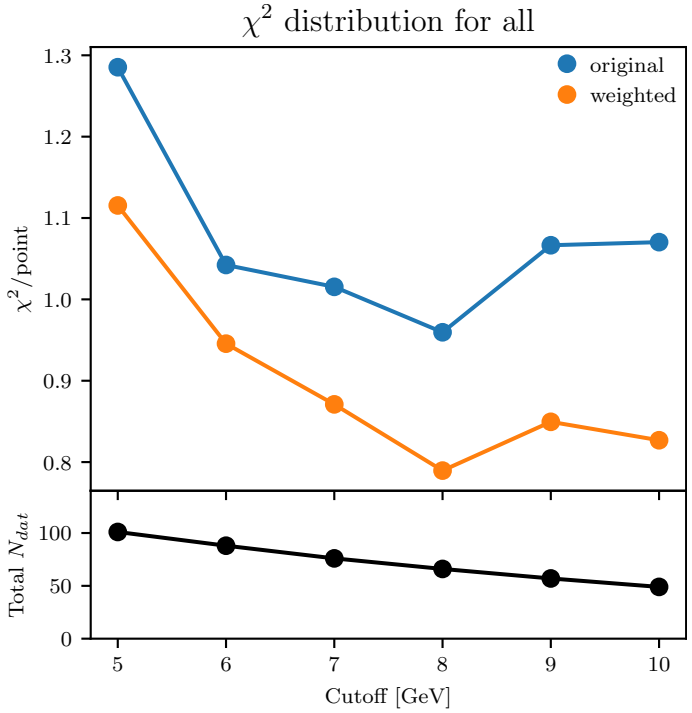

Figure 5.7: Above: Variation of χ^2/point based on the choice of cutoff $p_T^{\min}$. Below: Number of points remaining for the given cutoff.

5.5 Implications

A primary motivation for this analysis was to see if results from Ref. [9] were consistent with the `MAPFF` set. In their paper, the most accurate sets used in pp-collisions were the Kretzer [39] and DSS [10] fragmentation function sets, where their main difference was in the hardness of gluons at high-z values. Both sets failed to predict pp-collision data in the high energy and high-p_T zone. However, both these sets (and other less predictive sets tested in [9]) all used a functional analytic form, with DSS the most complex and Kretzer just power law scaling. Even without fitting against pp-collision data, `MAPFF` produced consistent predictions. The functional form FF sets significantly differ from neural network sets, so the later are capable of varying more to fit experimental data. Therefore some of the authors' observations, such as soft high-z gluons better predicting pp-collision data, are not necessarily applicable to neural network fits. Rather, the neural network sets significantly improve over the functional form sets, indicating that the FF behavior is more complicated in the interpolation region $0 < z < 1$.

The disagreement of the pp-collision data and the functional FF sets is most apparent in the high-p_T region, which is not strongly sampled by the reweighted fit. A full test of the reweighted set's data-theory agreement requires sampling this region. Another important discrepancy is the lack of error estimation for the Kretzer set, and the large variance of the DSS set. The reweighted FF set provides the most constrained error estimation of the fit based on all three major semi-inclusive processes, and therefore may quantify whether or or not it agrees with high energy and high-p_T data.

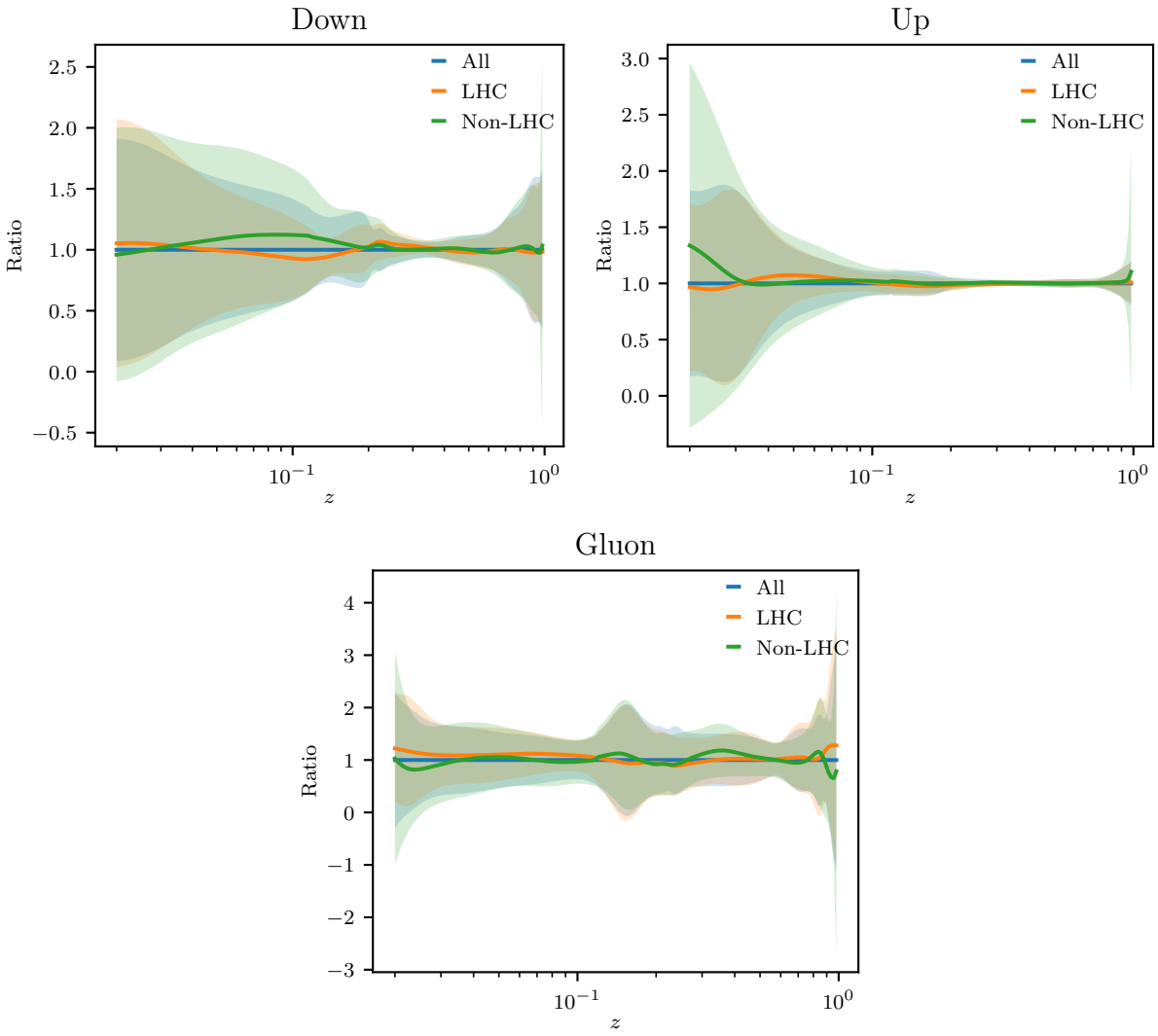

Figure 5.8: Sample ratios of different experimental sets against the total experimental set.

5.6 Improvements

The most substantive limiting factor in the analysis of pp-collisions is the lack of sampling the high energy and high-p_T domain, as previously stated. While data for the pion fragmentation function is limited in this domain, a similar analysis may be performed on other hadrons which admit a larger data range. Repeating this analysis for kaon, (anti-)protons, and unidentified charged hadrons, even on the same kinematic domain, provides a strong test of the results for the pion FF reweighting.

Similarly, a lack of identified charged pion data meant that the FF set could not be reweighted against charged pions individually, but rather against their average, i.e. summed and neutral pions. Only one type of charged pion needs to be collected in order to determine the other charge from the summed fit. However, this analysis is again limited by available data.

Available hadron-hadron collisions were omitted from this analysis due to time constraints. Collisions involving nuclei (i.e. Pb-Pb collisions) introduce nuclear effects and require more detailed corrections. However, this is still an artificial omission, and reweighting against these data sets will improve the quality of the FF fit.

Chapter 6

Conclusion

In this paper, the specific methods to link theoretical calculations, experimental data, and numeric computations of distribution functions are discussed, so that a researcher unfamiliar with the material may understand the internal behavior of modern PDF and FF programs (and a dedicated researcher may imagine how to produce such programs themselves). Emphasis is placed upon fragmentation functions, which are the subject of analysis for reweighting. However, it is necessary to identify their similarities and differences with the more abundant parton distribution functions, as well as the specifics of their treatment in fitting. PDFs have been successfully fitted using neural networks spanning multiple versions. FFs have only been recently fitted using neural networks, and this is the first time charged pion FFs have been fitted against semi-inclusive hadron-hadron collision data as well as the first time being reweighted. Still, PDF analysis serves as a baseline to compare this FF analysis.

Reweighting showed to be a useful technique for incorporating new data into a previously fitted FF set. As expected, it improved the fit against the total experimental data, and tended to improve individual fits for each ensemble replica. Unweighting the reweighted ensemble was successful in reproducing the probability distribution of the ensemble while reducing the ensemble size to the one predicted by Shannon information entropy.

Surprisingly, the original SIA and SIDIS fit was capable of reproducing most of the pp-collision data, including in the gluon partonic dominance range. Instead, the quark FFs received the greatest constraints under the new data. The high energy (> 1 TeV) and high-p_T ($\gtrsim 20$ GeV) range is the only part to be noticeably affected through reweighting, but is under-sampled in the experimental data. Future fittings should focus on data in the region to test whether hadron-hadron data does actually constrain FFs here.

Finally, the determination of FFs of the light hadrons is necessary to give accurate determinations of other topics of research, such as transverse-momentum-dependent distributions (TMDs) [8]. Parton in this paper are assumed to be collinear with the hadron, but in general they may carry a non-vanishing momentum transverse to the hadron. TMD FFs rely upon collinear FFs, so the accurate determination of collinear FFs will improve the fit of TMD FFs.

Appendix A

Distribution Function Distances

The distance between distribution functions is defined for reweighted distribution functions in [15]. The measurement is defined to determine if the same underlying distribution (i.e. the true physical distribution) generates both sets. A distance below approximately 2.3 at a point indicates the functions are generated by the same underlying distribution with about 90% confidence. The same applies to the variance of the distributions. Unlike the methods in the reference, the averages were not taken from random subsets of the ensemble since the `MAPFF` ensemble was twice as large as the `NNPDF` ensemble. Figure A.1 shows the distance between the original and reweighted FFs, as well as the distance between their uncertainties. The vast majority of the distribution functions lie within the 90% confidence interval, and what parts do no lie close to the boundary. Thus the original and reweighted distributions are very likely generated by the same underlying physical distribution.

Figure A.2 is the same analysis except between the reweighted and unweighted distributions. Again for the same reasons, the distributions are very likely generated by the same underlying physical distribution.

Because the corresponding ratio plots of the above comparisons show the central values are well encapsulated in the variance, the fact they're generated by the same underlying distribution may be concluded directly from the plots. The analysis of distances is included here for completeness.

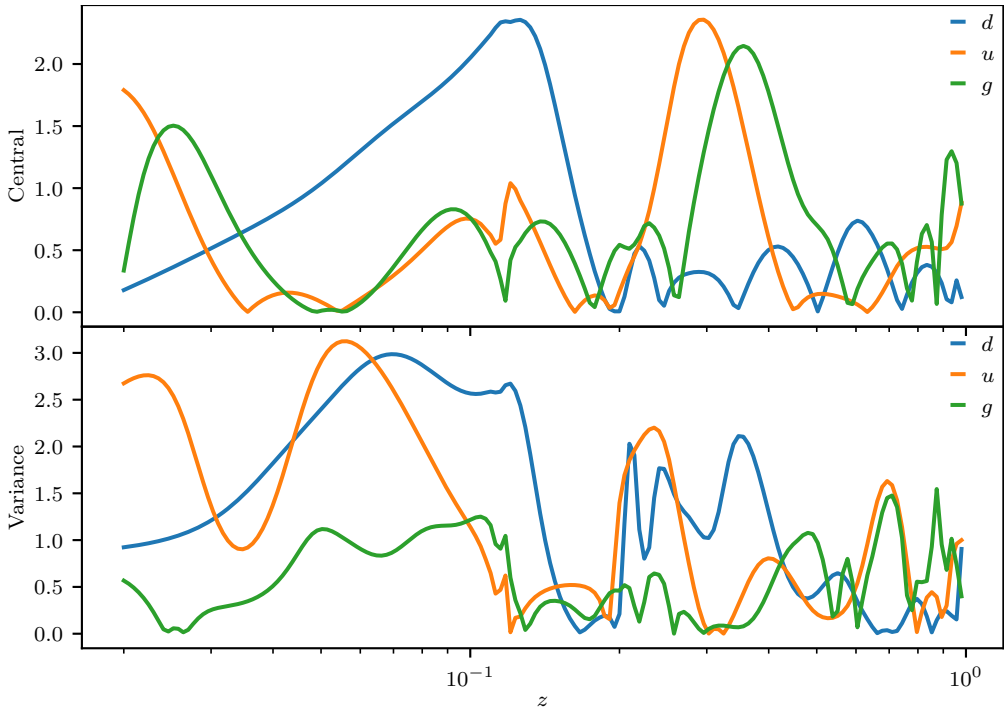

Figure A.1: Distance of original and reweighted FFs and their uncertainties at $\mu = 5GeV$.

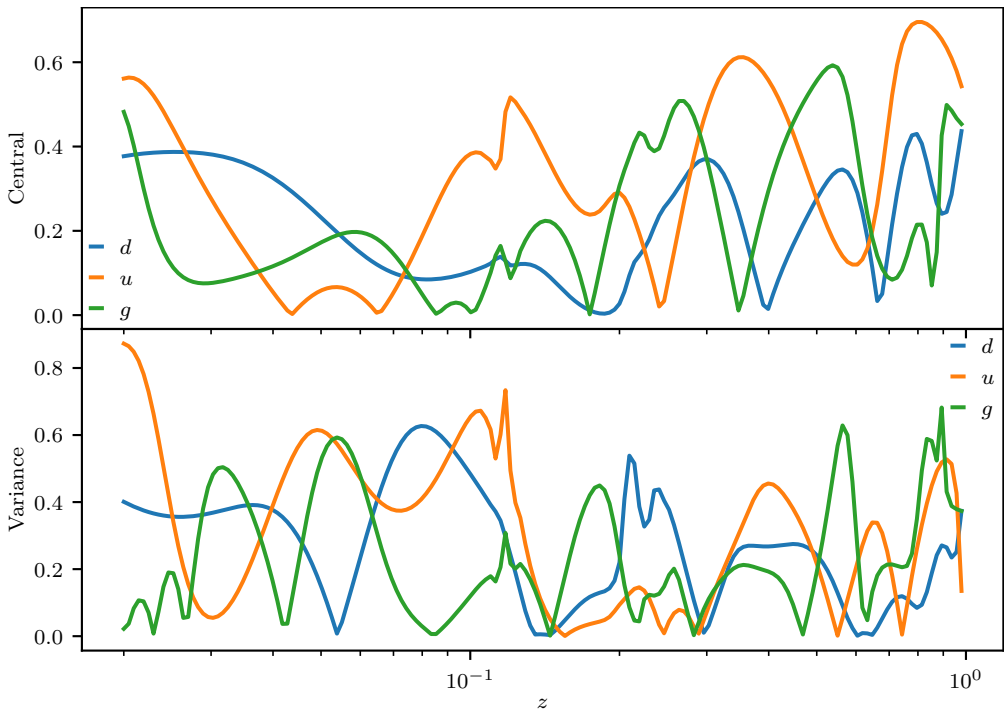

Figure A.2: Distance between weighted and unweighted distributions at $\mu = 5GeV$.

Appendix B

χ^2 Scale Estimation

The probability density $\mathcal{P}(\alpha)$ is a measure of consistency of the reweighting [15], where α rescales the replica χ^2_k values by χ^2_k/α^2. Reweighting by this new χ^2_k value for each replica quantifies the underestimation of uncertainty in the χ^2 analysis. A peak near one, as shown in B.1, indicates the uncertainties are well estimate by the reweighting.

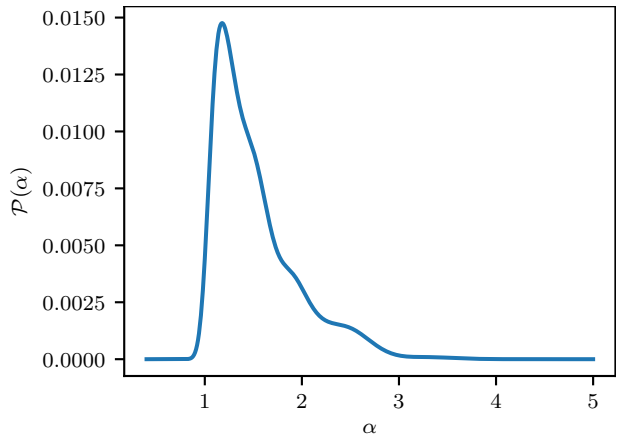

Figure B.1: Probability density of rescaling parameter α.

Appendix C

Code

All code, excluding external libraries, is documented and made available at the following GitHub repository:

`https://github.com/Hollenbeck-Hayden/MSc_colliding_partons.git`

Build instructions, including external dependencies, is listed in the repository documentation. An instance of the data generated and used in this report is provided, including an unweighted fragmentation function set in an LHAPDF compatible format. Instructions for how to include new data sets are also provided.

Bibliography

[1] Andy Buckley, James Ferrando, Stephen Lloyd, Karl Nordström, Ben Page, Martin Rüfenacht, Marek Schönherr, and Graeme Watt. Lhapdf6: parton density access in the lhc precision era. *The European Physical Journal C*, 75(3), Mar 2015.

[2] Valerio Bertone, Stefano Carrazza, and Juan Rojo. Apfel: A pdf evolution library with qed corrections. *Computer Physics Communications*, 185(6):1647–1668, Jun 2014.

[3] Valerio Bertone. Apfel++: A new pdf evolution library in c++, 2017.

[4] Valerio Bertone, Stefano Carrazza, Nathan P. Hartland, Emanuele R. Nocera, and Juan Rojo. A determination of the fragmentation functions of pions, kaons, and protons with faithful uncertainties. *The European Physical Journal C*, 77(8), Aug 2017.

[5] Rabah Abdul Khalek, Valerio Bertone, and Emanuele R. Nocera. Determination of unpolarized pion fragmentation functions using semi-inclusive deep-inelastic-scattering data. *Physical Review D*, 104(3), Aug 2021.

[6] J. C. Collins, D. E. Soper, and G. Sterman. Factorization of hard processes in qcd. *Adv. Ser. Direct. High Energy Phys.*, 5, 1988.

[7] Alberto Accardi, Francois Arleo, William Brooks, David D'Enterria, and Valeria Muccifora. Parton propagation and fragmentation in qcd matter. *Rivista Del Nueovo Cimento*, Jul 2009.

[8] A. Metz and A. Vossen. Parton fragmentation functions. *Progress in Particle and Nuclear Physics*, 91:136–202, Nov 2016.

[9] David d' Enterria, Kari J. Eskola, Ilkka Helenius, and Hannu Paukkunen. Confronting current nlo parton fragmentation functions with inclusive charged-particle spectra at hadron colliders. *Nuclear Physics B*, 883:615–628, Jun 2014.

[10] Daniel de Florian, Rodolfo Sassot, and Marco Stratmann. Global analysis of fragmentation functions for pions and kaons and their uncertainties. *Physical Review D*, 75(11), Jun 2007.

[11] Richard D. Ball, Valerio Bertone, Stefano Carrazza, Luigi Del Debbio, Stefano Forte, Patrick Groth-Merrild, Alberto Guffanti, Nathan P. Hartland, Zahari Kassabov,

José I. Latorre, and et al. Parton distributions from high-precision collider data. *The European Physical Journal C*, 77(10), Oct 2017.

[12] Kurt Hornik. Approximation capabilities of multilayer feedforward networks. *Neural Networks*, 4(2):251–257, 1991.

[13] J. Pumplin, D. Stump, R. Brock, D. Casey, J. Huston, J. Kalk, H. L. Lai, and W. K. Tung. Uncertainties of predictions from parton distribution functions. ii. the hessian method. *Physical Review D*, 65(1), Dec 2001.

[14] Stefano Forte, Lluí s Garrido, José I Latorre, and Andrea Piccione. Neural network parametrization of deep-inelastic structure functions. *Journal of High Energy Physics*, 2002(05):062–062, May 2002.

[15] Richard D. Ball, Valerio Bertone, Francesco Cerutti, Luigi Del Debbio, Stefano Forte, Alberto Guffanti, José I. Latorre, Juan Rojo, and Maria Ubiali. Reweighting nnpdfs: The w lepton asymmetry. *Nuclear Physics B*, 849(1):112–143, Aug 2011.

[16] The NNPDF Collaboration, Valerio Bertone, Nathan P. Hartland, Emanuele R. Nocera, Juan Rojo, and Luca Rottoli. Charged hadron fragmentation functions from collider data, 2019.

[17] R.K. Ellis W.J. Stirling and B.R. Webber. *QCD and Collider Physics*. Cambridge University Press, 1996.

[18] R.G. Roberts. *The Structure of the Proton*. Cambridge University Press, 1990.

[19] Kalanand Mishra, Luca Barzè, Thomas Becher, Mauro Chiesa, Stefan Dittmaier, Xavier Garcia i Tormo, Alexander Huss, Tobias Kasprzik, Ye Li, Guido Montagna, Mauro Moretti, Oreste Nicrosini, Frank Petriello, Fulvio Piccinini, and Francesco Tramontano. Electroweak corrections at high energies, 2013.

[20] J. P. Lees, V. Poireau, V. Tisserand, E. Grauges, A. Palano, G. Eigen, B. Stugu, D. N. Brown, L. T. Kerth, Yu. G. Kolomensky, and et al. Production of charged pions, kaons, and protons in e^+e^- annihilations into hadrons at $\sqrt{s} = 10.54$ gev. *Physical Review D*, 88(3), Aug 2013.

[21] Valerio Bertone, Stefano Carrazza, and Emanuele R. Nocera. Reference results for time-like evolution up to $o(\alpha_s^3)$. *Journal of High Energy Physics*, 2015(3), Mar 2015.

[22] C. Adolph, M. Aghasyan, R. Akhunzyanov, G.D. Alexeev, M.G. Alexeev, A. Amoroso, V. Andrieux, N.V. Anfimov, V. Anosov, K. Augsten, and et al. Multiplicities of charged pions and charged hadrons from deep-inelastic scattering of muons off an isoscalar target. *Physics Letters B*, 764:1–10, Jan 2017.

[23] K. Aamodt, N. Abel, U. Abeysekara, A. Abrahantes Quintana, A. Abramyan, D. Adamová, M. M. Aggarwal, G. Aglieri Rinella, A. G. Agocs, and et al. Production of pions, kaons and protons in pp collisions at $\sqrt{s} = 900$GeV with alice at the lhc. *The European Physical Journal C*, 71(6), Jun 2011.

[24] B. Abelev, J. Adam, D. Adamová, M. M. Aggarwal, M. Agnello, A. Agostinelli, N. Agrawal, Z. Ahammed, N. Ahmad, and et al. Neutral pion production at midrapidity in pp and pb–pb collisions at $\sqrt{s_{\mathrm{NN}}} = 2.76\,\mathrm{tev}$. *The European Physical Journal C*, 74(10), Oct 2014.

[25] S. Acharya, D. Adamová, M. M. Aggarwal, G. Aglieri Rinella, M. Agnello, N. Agrawal, Z. Ahammed, N. Ahmad, S. U. Ahn, and et al. Production of π^0 and η mesons up to high transverse momentum in pp collisions at 2.76 tev. *The European Physical Journal C*, 77(5), May 2017.

[26] B. Abelev, J. Adam, D. Adamová, M.M. Aggarwal, G. Aglieri Rinella, M. Agnello, A. Agostinelli, N. Agrawal, Z. Ahammed, N. Ahmad, and et al. Production of charged pions, kaons and protons at large transverse momenta in pp and pb–pb collisions at $\sqrt{s}_{\mathrm{nn}} = 2.76\mathrm{TeV}$. *Physics Letters B*, 736:196–207, Sep 2014.

[27] S. Acharya, D. Adamová, S. P. Adhya, A. Adler, J. Adolfsson, M. M. Aggarwal, G. Aglieri Rinella, M. Agnello, N. Agrawal, Z. Ahammed, and et al. Production of charged pions, kaons, and (anti-)protons in pb-pb and inelastic pp collisions at $\sqrt{s}_{\mathrm{nn}} = 5.02\mathrm{TeV}$. *Physical Review C*, 101(4), Apr 2020.

[28] B. Abelev, A. Abrahantes Quintana, D. Adamová, A.M. Adare, M.M. Aggarwal, G. Aglieri Rinella, A.G. Agocs, A. Agostinelli, S. Aguilar Salazar, Z. Ahammed, and et al. Neutral pion and η meson production in proton–proton collisions at $\sqrt{s} = 0.9\mathrm{TeV}$ and $\sqrt{s} = 7000\mathrm{TeV}$. *Physics Letters B*, 717(1-3):162–172, Oct 2012.

[29] J. Adam, D. Adamová, M. M. Aggarwal, G. Aglieri Rinella, M. Agnello, N. Agrawal, Z. Ahammed, I. Ahmed, S. U. Ahn, and et al. Measurement of pion, kaon and proton production in proton–proton collisions at $\sqrt{s} = 7\mathrm{TeV}$. *The European Physical Journal C*, 75(5), May 2015.

[30] S. Acharya, J. Adam, D. Adamová, J. Adolfsson, M. M. Aggarwal, G. Aglieri Rinella, M. Agnello, N. Agrawal, Z. Ahammed, and et al. π^0 and η meson production in proton-proton collisions at $\sqrt{s} = 8\mathrm{TeV}$. *The European Physical Journal C*, 78(3), Mar 2018.

[31] A. Adare, S. Afanasiev, C. Aidala, N. N. Ajitanand, Y. Akiba, H. Al-Bataineh, J. Alexander, K. Aoki, L. Aphecetche, R. Armendariz, and et al. Identified charged hadron production in p+pcollisions at $\sqrt{s} = 200$ and $62.4\mathrm{GeV}$. *Physical Review C*, 83(6), Jun 2011.

[32] A. Adare, C. Aidala, N. N. Ajitanand, Y. Akiba, R. Akimoto, H. Al-Ta'ani, J. Alexander, K. R. Andrews, A. Angerami, K. Aoki, and et al. Charged-pion cross sections and double-helicity asymmetries in polarized $p + p$ collisions at $\sqrt{s} = 200$ GeV. *Phys. Rev. D*, 91:032001, Feb 2015.

[33] S. S. Adler, S. Afanasiev, C. Aidala, N. N. Ajitanand, Y. Akiba, J. Alexander, R. Amirikas, L. Aphecetche, S. H. Aronson, R. Averbeck, and et al. Midrapidity neutral-pion production in proton-proton collisions at $\sqrt{s} = 200\mathrm{GeV}$. *Physical Review Letters*, 91(24), Dec 2003.

[34] A. Adare, C. Aidala, N. N. Ajitanand, Y. Akiba, R. Akimoto, J. Alexander, M. Alfred, K. Aoki, N. Apadula, Y. Aramaki, and et al. Inclusive cross section and double-helicity asymmetry for π^0 production at midrapidity inp+pcollisions at $\sqrt{s} = 510$GeV. *Physical Review D*, 93(1), Jan 2016.

[35] S. Chatrchyan, V. Khachatryan, A. M. Sirunyan, A. Tumasyan, W. Adam, E. Aguilo, T. Bergauer, M. Dragicevic, J. Erö, and et al. Study of the inclusive production of charged pions, kaons, and protons in pp collisions at $\sqrt{s} = 0.9, 2.76,$ and 7 TeV. *The European Physical Journal C*, 72(10), Oct 2012.

[36] B. I. Abelev, M. M. Aggarwal, Z. Ahammed, A. V. Alakhverdyants, B. D. Anderson, D. Arkhipkin, G. S. Averichev, J. Balewski, O. Barannikova, L. S. Barnby, and et al. Longitudinal double-spin asymmetry and cross section for inclusive neutral pion production at midrapidity in polarized proton collisions at $\sqrt{s} = 200$GeV. *Physical Review D*, 80(11), Dec 2009.

[37] B. I. Abelev, M. M. Aggarwal, Z. Ahammed, A. V. Alakhverdyants, B. D. Anderson, D. Arkhipkin, G. S. Averichev, J. Balewski, O. Barannikova, L. S. Barnby, and et al. Inclusive π^0, η and direct photon production at high transverse momentum inp+pandd+aucollisions at $\sqrt{s}_{\rm nn} = 200$GeV. *Physical Review C*, 81(6), Jun 2010.

[38] Richard D. Ball, Luigi Del Debbio, Stefano Forte, Alberto Guffanti, José I. Latorre, Juan Rojo, and Maria Ubiali. Fitting parton distribution data with multiplicative normalization uncertainties. *Journal of High Energy Physics*, 2010(5), May 2010.

[39] S. Kretzer. Fragmentation functions from flavor-inclusive and flavor-tagged e^+e^- annihilations. *Physical Review D*, 62(5), Jul 2000.

[40] Diederik P. Kingma and Jimmy Ba. Adam: A method for stochastic optimization, 2017.

[41] NNPDF Collaboration, Richard D. Ball, Valerio Bertone, Francesco Cerutti, Luigi Del Debbio, Stefano Forte, Alberto Guffanti, Nathan P. Hartland, Jose I. Latorre, Juan Rojo, and Maria Ubiali. Reweighting and unweighting of parton distributions and the lhc w lepton asymmetry data, 2011.

[42] Daniel de Florian. Next-to-leading order QCD corrections to one hadron production in polarized pp collisions at RHIC. *Phys. Rev. D*, 67:054004, 2003.

[43] B. Jager, A. Schafer, M. Stratmann, and W. Vogelsang. Next-to-leading order QCD corrections to high p(T) pion production in longitudinally polarized pp collisions. *Phys. Rev. D*, 67:054005, 2003.